

Introduction to Formal Reasoning and Decision Making

Juan Comesaña

October 24, 2026

Preface

This is a textbook to accompany Philosophy 109 at Rutgers New Brunswick. I have borrowed extensively from teaching notes by Max Bialek, Brian Weatherson, and Jonathan Weisberg.

Contents

Contents	ii
Part I Logic	1
1 The language of Propositional Logic	2
1.1 Introduction	2
1.2 Sentences and propositions	3
1.3 Atomic vs. molecular propositions	4
1.4 The syntax of PL	6
1.4.1 Syntax trees and conventions	8
1.5 The Semantics of PL: truth-tables	11
1.5.1 Truth-functional completeness and disjunctive normal form	15
2 Translations	21
2.1 From English to PL	21
2.2 Translation notes for connectives	23

2.2.1	Negation	24
2.2.2	Conjunction and its English cousins	24
2.2.3	Disjunction—inclusive and exclusive	25
2.2.4	Conditional	26
2.2.5	Biconditional	27
2.2.6	Some other expressions	28
3	Arguments	30
3.1	Good Arguments	30
3.2	Validity	31
3.2.1	Validity, soundness and good arguments	35
3.3	Tautological Validity	37
3.4	Translating arguments	40
Part II	Probability	44
4	Measures and stochastic truth-tables	45
4.1	Introduction	45
4.2	Measures	46
4.2.1	Normalized measures	48
4.2.2	Possibility space	48
4.2.3	Normalized measures over possibility space and stochastic truth-tables	49
4.3	Conditional probability	50
5	Axioms, definitions, and theorems	54
5.1	Axioms	54
5.2	Independence	55
5.3	Unconditional probability theorems	56
5.4	Conditional probability theorems	60
5.5	Solving problems	62
6	Bayesianism	68
6.1	The lottery paradox: beliefs and credences	68
6.2	Bayesianism	69
6.3	Eikosograms	77
6.4	Hume on miracles	79

6.5	The problem of priors	85
Part III Decision and Game Theory		91
7	Decisions	92
7.1	Representing decisions	92
7.2	Where do ordinal utilities come from?	95
7.3	Choice under ignorance	98
7.3.1	Dominance	99
7.3.2	Maximin	103
7.3.3	Minimax regret	105
8	Decisions under risk	111
8.1	Cardinal utilities	111
8.2	Expected value	114
8.3	Where do cardinal utilities come from?	115
8.4	Maximization of expected utility	116
8.5	When states depend on choices: maximization of conditional expected utility	119
9	Newcomb's problem	124
9.1	Introduction	124
9.2	The case	124
9.3	The initial arguments	128
9.4	"Why ain'cha rich," lack of opportunities, and transparent Newcomb	130
10	Games	136
10.1	Games as interactive decisions	136
10.1.1	Matrices	137
10.1.2	Trees	138
10.1.3	Example: Rock-Paper-Scissors	139
10.1.4	From trees to matrices, and back	141
10.2	The assumptions of game theory	143
10.3	Strictly competitive two-player games	144
10.4	Nonzero-sum games	145
10.5	Iterated deletion of dominated strategies	146
10.6	Nash equilibrium	151

10.6.1	Saddle points	156
10.6.2	Finding pure equilibria by starring best responses	158
10.7	The Prisoner's Dilemma	160
10.7.1	The case	160
10.7.2	Pareto optimality	162
10.7.3	Two mistaken reactions	162
10.7.4	The Prisoner's Dilemma and Newcomb's problem	164
10.8	Backwards Induction	166
10.8.1	Playing the Prisoner's Dilemma more than once	166
10.8.2	The backwards induction argument	167
10.8.3	The centipede game	168
10.8.4	The role of common knowledge	169
Appendix: Solutions to the Exercises		177
Chapter 1		177
Chapter 2		178
Chapter 3		179
Chapter 4		179
Chapter 5		180
Chapter 6		180
Chapter 7		181
Chapter 8		181
Chapter 9		181
Chapter 10		182
Appendix: Formula Sheet		184
Appendix: Fraction Rules		185

Part I

Logic

1 The language of Propositional Logic

This chapter introduces the basics of propositional logic and the method of truth-tables.

1.1 Introduction

Suppose you don't know whether "Leona" is the name of an animal or the name of the statue decorating the entrance to my house. I then tell you that Leona is a cat, and that if Leona is a cat then Leona is an animal. Because you trust me, you come to believe that Leona is an animal on the basis of my claims that Leona is a cat and that if Leona is a cat then Leona is an animal (you didn't need me to tell you that this last claim is true, obviously). We can reconstruct what I did in the form of an *argument* (for now, this is just an example of an argument—we'll give a more formal treatment of that notion later on):

- (1) Leona is a cat.
- (2) If Leona is a cat, then Leona is an animal.

∴ Leona is an animal.

In the argument above, the enumerated sentences are its *premises*, and the "∴" symbol (read as "therefore") precedes the *conclusion*. That argument has an interesting feature: it's impossible for its premises to be true and its conclusion false. By contrast, consider the following argument:

- (1) Leona is an animal.
- (2) If Leona is a cat, then Leona is an animal.

$\therefore$ Leona is a cat.

This second argument does not have the feature we identified in the first one: it is possible for its premises to be true and its conclusion false. Our aim in this first part is to give a mathematical model of that feature in question, which logicians call *validity*. To say that an argument is valid is not the same as saying that it is a *good* argument. Some valid arguments are really bad (we'll see examples later). And some very good arguments are not valid. Nevertheless, the notion of validity plays an important role in the evaluation of some arguments, so we will start by analyzing it. Our strategy to analyze the notion of validity will be to construct an *artificial language*, the language of *propositional logic* (**PL**). Inside **PL** we can then construct a toy model of the notion of validity, *tautological validity*.

1.2 Sentences and propositions

Unsurprisingly, *propositions* will be the main elements of **PL**. There is a tight connection between propositions and some sentences, but there are also important differences.

A proposition is a bit of information, something which can be true or false. Thus, that Leona is a cat is a proposition, and so is the proposition that Leona is an animal. It is either true or false that Leona is a cat, and also either true or false that Leona is an animal. Whether you know if Leona is a cat or an animal does not affect the fact that those are indeed propositions, for the defining characteristic of a proposition is that it can be true or false, not that anyone knows whether it is true or false.

"Leona is a cat" is a sentence of English. What is the relationship between that sentence and the proposition that Leona is a cat? Could we perhaps identify them? Sentences, after all, can be true or false (at least some of them, anyway, and certainly "Leona is a cat"). No, we will not identify sentences and propositions. Rather, we will say that some sentences *express* a proposition. For instance, "Leona is a cat" expresses the proposition that Leona is a cat. But that very proposition can be expressed by other sentences. For instance, the sentence "Leona is a member of the species *felis catus*" expresses that same proposition. And so do the sentences "Leona es una gata" and "Leona est une chatte". Because all of those sentences are different from each other, we cannot identify any of them with the proposition that Leona is a cat. Rather, that proposition is what all those different sentences

express. "Leona es una gata" is a good translation to Spanish of "Leona is a cat" precisely because both sentences express the same proposition.¹

Just as different sentences can express the same proposition, so too the same sentence can express different propositions. For instance, the sentence "I am hungry" expresses one proposition when I say it and a different one when you say it. So that is another reason why we cannot identify propositions with sentences.

But there is still a tight connection between sentences and propositions, because we said that sentences can *express* propositions. Another way of saying the same thing is that propositions can be the meaning, or the content, or that which is said by, some sentences. The qualifier "some" is important here, because not all sentences express propositions. Sentences in the imperative and interrogative mood usually do not ("Close the door" expresses a command, and "Is the door closed?" a question, but neither commands nor questions can be true or false). Should we say, then, that a proposition is the meaning of a sentence in the indicative mood? Not really, because some sentences in the indicative mood do not have a content which can be true or false—or at least that is what some philosophers would say. Suppose, for instance, that I say "That is green" while pointing at nothing in particular. I have not managed to express any truth-evaluable content, but yet I have uttered a sentence in the indicative mood. For a different kind of example take the sentence "I apologize". According to some philosophers, what that sentence expresses is not a proposition that can be true or false—in particular, it does not express the proposition that the speaker apologizes. Rather, in uttering that sentence the speaker is apologizing.²

So what can we say, then, about the relationship between sentences and propositions? We can say that a sentence expresses a proposition when its content (or its meaning, or what the sentence says) is something that can be true or false. Usually (but not always) sentences in the indicative mood will be used to convey information, and thus will express propositions.

1.3 Atomic vs. molecular propositions

Some propositions have other propositions as parts. Consider, for instance, the proposition:

¹I'm ignoring here the complications having to do with the fact that Spanish and French are grammatically gendered languages, whereas English is not.

²According to these philosophers, in saying "I apologize" under the right conditions, I perform a *speech act*, in this case the act of apologizing.

Leona is a cat, and she is beautiful

Inside that proposition we can find these two propositions:

Leona is a cat.

Leona is beautiful.

For that reason, we will say that the proposition *Leona is a cat, and she is beautiful* is a *molecular* proposition. By contrast, consider the proposition *Leona is a cat*. That proposition does arguably have parts—for instance, one part corresponding to Leona, and another one referring to the class of cats. But those parts are not themselves propositions. We will call propositions that do not have other propositions as parts *atomic* propositions. Thus, just as chemical molecules have atoms as parts but atoms do not themselves have atoms as parts (although they do have other parts, such as electrons), so too molecular propositions have other propositions as parts but atomic propositions do not have other propositions as parts (although they do have other parts).

In chemistry, the "glue" that binds atoms together into molecules are electromagnetic interactions. What is the glue that binds atomic propositions into molecular ones? They are *truth-functional connectives*. In the case of *Leona is a cat, and she is beautiful*, the truth-functional connective is the particle "and"—known as *conjunction* in logic. It is a truth-functional connective because we can think of it as a function that takes truth-values as arguments and delivers truth-values as values. The definition of conjunction is that it delivers a true proposition when applied to two true propositions as arguments, and a false proposition in every other case. Thus, because it is true that Leona is a cat and is also true that Leona is beautiful, it is true that *Leona is a cat, and she is beautiful*.

The other truth-functional connectives that we will work with are: negation, (inclusive) disjunction, material conditional, and material biconditional. In this chapter we will define those truth-functional operations both syntactically and semantically in the context of our language of propositional logic, and in chapter 2 we will study the connections between that language and natural languages such as English.

1.4 The syntax of PL

So far we have been using sentences in English to express propositions. **PL**, however, is not English. It has an indefinite number of *atomic sentence letters*, which we write as upper-case letters of the alphabet, with indices if needed. Thus, the following are all atomic sentence letters of *Propositional Logic* (in what follows, we abbreviate the language as **PL**): $A, B, C, \dots, P_1, S_{323}, \dots$. Such a letter can be used to stand for an atomic proposition, but it is a symbol of **PL**, not a proposition. For instance, if we let L stand for the proposition that Leona is a cat and B for the proposition that Leona is beautiful, then the English sentence *Leona is a cat, and she is beautiful* could be translated into **PL** as $L \wedge B$.

As we said, we will work with five truth-functional connectives: negation, conjunction, disjunction, material conditional, and material biconditional. The symbols we will use for those connectives are as follows:

Truth-functional connectives

Negation: $\neg$

Conjunction: $\wedge$

Disjunction: $\vee$

Conditional: $\rightarrow$

Biconditional: $\leftrightarrow$

So far we have truth-functional connectives and atomic propositions as part of **PL**. But we need one more component. For consider the string of symbols $L \wedge B \vee C$. Is that supposed to be a conjunction of L with $B \vee C$ or a disjunction of $L \wedge B$ with C ? That string of symbols is ambiguous between those two meanings, so we need a way to syntactically distinguish between them. Parentheses to the rescue: we can distinguish them by writing $L \wedge (B \vee C)$ or $(L \wedge B) \vee C$. That completes our language of **PL**:

Language of Propositional Logic

Atomic sentence letters: $A, B, C, \dots, A_1, B_1, C_1, \dots$

Truth-functional connectives: $\neg, \wedge, \vee, \rightarrow, \leftrightarrow$.

Parentheses: $(,)$.

We will soon start writing sentences of **PL** and analyzing their structure. To do so, however, we need to *talk about* sentences of **PL**. **PL** itself is not complex enough to allow us to use *it* to talk about itself. So instead we will use English to talk about **PL**, just as we have been doing so far. For instance, the sentence "Parentheses are part of the vocabulary of **PL**" is an English sentence about **PL**. But we are going to augment English with some math vocabulary in order to be able to state general truths about **PL**. For instance, we want to distinguish between those sentences of **PL** that are grammatically correct (we'll call them well-formed sentences, or *wffs* for short) from those that are not. We can point at instances of that distinction—for instance, we can say that whereas $A \vee B$ is grammatically correct, $\vee AB$ is not. But we want to make general claims about wffs, not just point to instances. It won't do to try to generalize by saying something like the following: "for every A and B , $A \vee B$ is a wff but $\vee AB$ is not". In that sentence, we are trying to use A and B as variables that range over every wff of **PL**, but A and B are not variables that range over sentences of **PL**, they are themselves two sentences of **PL**. We could get by by saying the following: "if you take any two wffs of **PL** and put a $\vee$ in between them, the result is a wff, whereas if you put a $\vee$ in front of them, the result is not a wff." But it is easier to understand if we use *meta-linguistic variables* to talk about arbitrary sentences of **PL**: "If φ and ψ are wffs, then so is $\varphi \vee \psi$, but $\vee \varphi \psi$ is not." We will use greek letters as our meta-linguistic variables (variables that we use to talk about **PL**).

We are now in a position to give a full syntax for **PL**. A syntax for a language is a specification of which combination of symbols of the language constitute grammatically correct sentences of that language (wffs).

Syntax of PL

Base clause

Every atomic sentence letter is a wff.

Recursive clause

If φ and ψ are wffs, then so are:

- $(\neg \varphi)$
- $(\varphi \vee \psi)$
- $(\varphi \wedge \psi)$
- $(\varphi \rightarrow \psi)$

- $(\varphi \leftrightarrow \psi)$

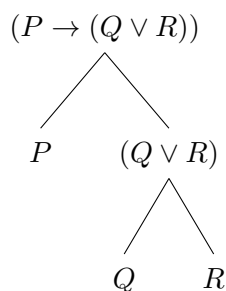
Closure clause

Nothing else is a wff.

The syntax of **PL** takes the form of a *recursive definition* of wff. One important thing to note about it is that its recursive clause seems to assume that we already know what a wff is. Isn't this viciously circular? Wasn't the definition supposed to tell us what a wff is? How can it then assume that we can already recognize whether something is a wff? The circularity is merely apparent. Given any finite combination of symbols of **PL**, the definition allows us to determine whether it is a wff or not. Consider, for instance, the formula $(P \rightarrow (Q \vee R))$. Notice that the recursive clauses connect at most two wffs to make up larger wffs (with the exception of the sub-clause for the negation connective, which applies to just one wff). In this case, we know that we are dealing with a formula of the form $\varphi \rightarrow \psi$.³ So, by the sub-clause of the recursive clause that deals with $\rightarrow$, we know that $(P \rightarrow (Q \vee R))$ is a wff just in case both P and $(Q \vee R)$ are wffs. Because P is an atomic sentence letter, it is a wff by the base clause. $(Q \vee R)$ has the form $(\varphi \vee \psi)$, and so it is a wff just in case both Q and R are wffs—and they are, because they are both atomic sentence letters. We can make this process systematic by using *syntax trees*.

1.4.1 Syntax trees and conventions

A syntax tree tells the story of how a wff is built up out of atomic sentence letters. For instance, the syntax tree for our example formula $(P \rightarrow (Q \vee R))$ is the following:



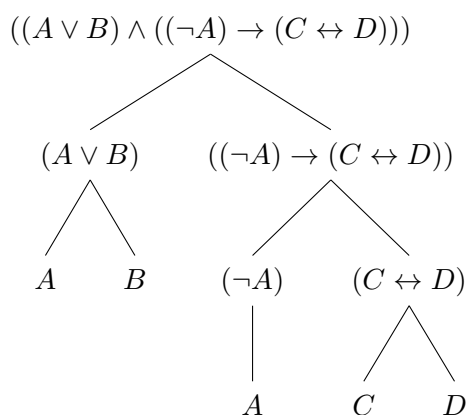
³How do we know that we are not dealing with a formula of the form $\varphi \vee \psi$? That is to say: how do we know that the *main connective* of the formula is $\rightarrow$ and not $\vee$? The short answer is that the main connective of a formula is the least embedded one. The long answer comes below, when we talk about *syntax trees*.

Try it in the Logic Lab ↗

Looking at this tree from top to bottom, we start with the complete formula and determine its main connective. The main connective of a wff is the connective whose *scope* is the whole formula. The *scope* of a connective is the set of formulas to which it applies. Thus, the main connective of a wff is *the* connective which applies to the whole formula (every wff that is not an atomic sentence letter has exactly one main connective; atomic sentence letters have none). In the formula above, $\vee$ cannot be the main connective because it does not apply to P —only $\rightarrow$ applies to both P and $(Q \vee R)$. In practical terms, the main connective will be the only connective enclosed only by the outermost parentheses of the formula. $\rightarrow$ connects P on the left to $(Q \vee R)$ on the right, so we draw two branches from the main formula and write down those two sub-formulae as nodes of the branches. When we reach an atomic sentence letter, that branch of the tree is complete. Thus, we no longer branch down from P . But, on the right-hand side, we have a new formula, $(Q \vee R)$. It is trivial in this case to determine the main connective of that formula, because it has only one: $\vee$. The connective, moreover, is connecting two atomic sentence letters. So, we branch down and terminate each branch with the corresponding atomic sentence letter.

We can also look at the tree from the bottom to the top, and that is how the tree tells the story of how the formula was constructed. We start out at the terminal nodes of the tree with three atomic sentence letters: P, Q , and R . We connect Q and R together with $\vee$, and then we connect the resulting formula, $(Q \vee R)$, to P with $\rightarrow$, resulting in the complete formula $(P \rightarrow (Q \vee R))$. Thus, the main connective is the last connective applied when constructing the formula.

Let's build a syntax tree for a more complex formula: $((A \vee B) \wedge ((\neg A) \rightarrow (C \leftrightarrow D)))$:



Try it in the Logic Lab ↗

This tree illustrates two important points. First, negation behaves differently in trees, because it is our only unary connective (all the others are binary connectives). Because negation applies to a single formula and not two, it generates just one branch. Second, a syntax tree can have multiple instances of the same atomic sentence letter as terminal nodes. In this case, A appears twice: once as an immediate component of $(\neg A)$ and once as an immediate component of $(A \vee B)$.

Syntax trees and our recursive definition for the syntax of **PL** go hand in hand: a formula is a wff of **PL** if and only if a syntax tree can be built up for it. Consider, for instance, the formula $(P \vee Q \wedge R)$. We cannot build a syntax tree for it because it has no unique main connective: both $\vee$ and $\wedge$ are enclosed only by the outermost parentheses (the only parentheses in the formula, in this case), and so there is no unique connective that is so enclosed. For the same reason, $(P \vee Q \wedge R)$ is not a wff according to our recursive definition, for it is neither an atomic sentence letter nor does it have any of the five forms specified in the recursive clause.

Let's go back to the formula $((A \vee B) \wedge ((\neg A) \rightarrow (C \leftrightarrow D)))$. It strictly adheres to the definition of wff, but it is a bit of an eyesore and not very easy to parse. We can adopt conventions to make formulae more readable. The first thing we can do is to eliminate the outermost parentheses. This gives us $(A \vee B) \wedge ((\neg A) \rightarrow (C \leftrightarrow D))$. Once we eliminate the outermost parentheses, we can no longer say that the main connective in a formula is the one that is enclosed only by the outermost parentheses—instead, the main connective in a formula is now the only one enclosed by no parentheses at all. We can also stipulate that negation applies to the wff that immediately follows it. Thus, although $\neg A \vee B$ is not strictly speaking a wff, it is equivalent to the following wff once we bring back the outermost parentheses and

enclose the negation sign and the wff that immediately follows it in parentheses as well: $((\neg A) \vee B)$. If instead we want the negation sign to apply to the disjunction $A \vee B$ we have to write $\neg(A \vee B)$. Using those two conventions together, the above formula becomes $(A \vee B) \wedge (\neg A \rightarrow (C \leftrightarrow D))$.

Parentheses-dropping conventions

- Outermost parentheses can be dropped.
- Negations have scope over the shortest wff following them.

1.5 The Semantics of PL: truth-tables

An atomic sentence letter can have one of two truth-values: it can be true, or it can be false. But what exactly does it mean, for instance, for “ A ” to be true? It depends on whether we are dealing with **uninterpreted** or **interpreted PL**. When we interpret **PL**, we make a dictionary to assign a natural-language atomic sentence to a **PL** atomic sentence letter. For instance, we can have the following assignment:

$$A = \text{Leona is a cat}$$

Under that interpretation, “ A ” is true just in case Leona is a cat. We will discuss interpretations of **PL** at length in the next chapter.

But if we have an uninterpreted sentence letter “ A ”, what would it mean for it to be true? Strictly speaking, nothing—it makes no sense to say that an uninterpreted sentence letter is true or false. But we can still suppose that “ A ” stands for an arbitrary (and unspecified) atomic proposition, which of course will be either true or false (this is why sometimes atomic sentence letters are called “propositional variables”). We can represent those two possibilities in a table:

A
T
F

If we now consider two atomic propositions at the same time, say A and B , then we have four possibilities in total:

In general, for n atomic propositions, there will be 2^n combinations of truth-values between them.

A	B
T	T
T	F
F	T
F	F

Let us now formally define our truth-functional connectives. We can do so by using the truth-tables just introduced:

Table 1.1: Conjunction

φ	ψ	$\varphi \wedge \psi$
T	T	T
T	F	F
F	T	F
F	F	F

Table 1.2: Disjunction

φ	ψ	$\varphi \vee \psi$
T	T	T
T	F	T
F	T	T
F	F	F

Table 1.3: Negation

φ	$\neg\varphi$
T	F
F	T

Table 1.4: Conditional

φ	ψ	$\varphi \rightarrow \psi$
T	T	T
T	F	F
F	T	T
F	F	T

Table 1.5: Biconditional

φ	ψ	$\varphi \leftrightarrow \psi$
T	T	T
T	F	F
F	T	F
F	F	T

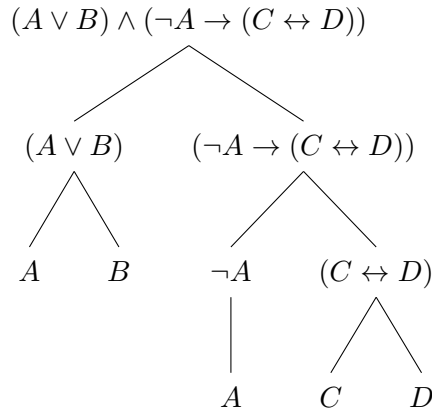
Notice that we defined the connectives using our Greek metalinguistic variables, which range over any wff. This is the right way to do it, because we want to define (e.g.) not just $A \wedge B$ but all wffs of arbitrary complexity whose main connective is conjunction.

The process of drawing the truth-table for an arbitrary wff of **PL** is algorithmic—you can program a computer to do it. Here is one way of specifying the algorithm:

Algorithm for constructing the truth-table for a given wff of PL:

1. To calculate the number of rows that you will need, count the number of distinct atomic sentence letters in the wff. For n distinct atomic sentence letters, your truth table will have 2^n rows of assignments. In addition, there will be one header row.
2. In the header row of the truth table, write down all the distinct atomic sentence letters in order of appearance in the wff, and then write down the wff itself.
3. For the i th atomic sentence letter, alternate between 2^{n-i} "T"s and 2^{n-i} "F"s until all rows are filled.
4. Create the syntax tree for the wff.
5. Use the syntax tree and the definitions of the connectives to recursively compute the truth-values for all the sub-formulae, including the full wff.

Step 5 is the crucial one. Consider again the following syntax tree:



Step 5 directs us to first calculate the truth-value of $\neg A$ using the definition of negation and the truth-values for A , $(C \leftrightarrow D)$ using the definition of biconditional and the truth-values for C and D , and $(A \vee B)$ using the definition of disjunction and the truth-values for A and B . These sub-formulae are all just one step above the terminal nodes of the tree. In the second step, we calculate the truth-value of $\neg A \rightarrow (C \leftrightarrow D)$ using the previously calculated values for $\neg A$ and $(C \leftrightarrow D)$. Finally, we calculate the truth-value of the whole formula, $(A \vee B) \wedge (\neg A \rightarrow (C \leftrightarrow D))$.

D)) using the definition of conjunction and the previously calculated values for $(A \vee B)$ and $\neg A \rightarrow (C \leftrightarrow D)$. The result is the following truth-table:

A	B	C	D	$(A \vee B) \wedge (\neg A \rightarrow (C \leftrightarrow D))$			
T	T	T	T	T	T	F	T
T	T	T	F	T	T	F	T
T	T	F	T	T	T	F	T
T	T	F	F	T	T	F	T
T	F	T	T	T	T	F	T
T	F	T	F	T	T	F	T
T	F	F	T	T	T	F	T
T	F	F	F	T	T	F	T
F	T	T	T	T	T	T	T
F	T	T	F	T	F	T	F
F	T	F	T	T	F	T	F
F	T	F	F	T	T	T	T
F	F	T	T	F	F	T	T
F	F	T	F	F	F	T	F
F	F	F	T	F	F	T	F
F	F	F	F	F	F	T	T

Try it in the Logic Lab ↗

What do the rows of a truth-table represent? They represent possibilities. Thus a row where both A and B are true represents a possibility where each of the propositions expressed by A and B are true. We can think of each row as a *possible world*: a way things could have been (as far as our atomic propositions are concerned).

A useful classification of wffs of **PL** distinguishes between *tautologies*, *contradictions*, and *contingencies*. A wff is a *tautology* when it is true in every row of its truth-table. It is a *contradiction* when it is false in every row of its truth-table. And it is a *contingency* when it is true in at least one row and false in at least one row. These notions are related to the notion of necessity. A *necessary* proposition is one that is true in *every* possible world, and an impossible proposition is one that is true in none. Since each row of a truth-table represents a possible world (distinguished no more finely than propositional logic allows), a tautology can only express a necessary proposition, and a contradiction can only express an impossible one.⁴ Here are examples of the three kinds of wffs:

⁴There are some necessary truths and impossibilities which cannot be recognized as such in

P	Q	$(P \rightarrow Q) \vee (Q \rightarrow P)$		
T	T	T	T	T
T	F	F	T	T
F	T	T	T	F
F	F	T	T	T

Try it in the Logic Lab ↗

A	B	$A \wedge (B \wedge \neg A)$	
T	T	F	F F
T	F	F	F F
F	T	F	T T
F	F	F	F T

Try it in the Logic Lab ↗

P	Q	$(P \wedge Q) \vee (P \wedge \neg Q)$		
T	T	T	T	F F
T	F	F	T	T T
F	T	F	F	F F
F	F	F	F	T F

Try it in the Logic Lab ↗

1.5.1 Truth-functional completeness and disjunctive normal form

We introduced five truth-functional connectives—a unary one (negation), and four binary ones (disjunction, conjunction, conditional and biconditional). A question you may be asking now is: why precisely those five? Why not more, or less?

Those are good questions! Something relevant to answering them is the issue of how many truth-functions there are. The question of how many truth-functions there are is actually incomplete, because it depends on how many variables we are talking about. A truth-function of n variables takes as input an assignment of truth-values to those variables, and delivers a truth-value as output. As we saw, there are 2^n such assignments (the rows of the truth-table). A truth-function is completely determined by which output, **T** or **F**, it delivers for each of those 2^n

propositional logic: for instance, that everything is self-identical is a necessary truth, and its negation an impossibility, but their representation in propositional logic makes them both come out as contingent.

inputs. Each of the 2^n rows can be given either of two outputs, independently of the others, and so the number of truth-functions of n variables is

$$\underbrace{2 \times 2 \times \cdots \times 2}_{2^n \text{ times}} = 2^{(2^n)}.$$

For the case of just one variable there are $2^1 = 2$ inputs (the variable can be true or false), and so $2^2 = 4$ truth-functions, represented in this table:

Table 1.6: The four truth-functions of one variable

φ	φ	$\neg\varphi$	$\varphi \vee \neg\varphi$	$\varphi \wedge \neg\varphi$
T	T	F	T	F
F	F	T	T	F

Notice that the first two functions are just φ and $\neg\varphi$. The third one is a tautological function. There are infinitely many ways of expressing a tautology with one variable: one is $\varphi \vee \neg\varphi$. Finally, the last one is a contradiction. There are also infinitely many ways of expressing a contradiction with one variable: one is $\varphi \wedge \neg\varphi$. This means that armed with just negation, disjunction and conjunction we can express all the possible truth-functions of one variable. So far, so good.

What happens when we move to truth-functions of two variables? Well, now we have $2^2 = 4$ inputs, and so $2^{2^2} = 16$ truth-functions. Will our five connectives be enough to represent them? Yes:

Table 1.7: The sixteen truth-functions of two variables

φ	ψ	$\varphi \wedge \neg\psi$	$\neg(\varphi \vee \psi)$	$\neg\varphi \wedge \psi$	$\neg\varphi$	$\varphi \wedge \neg\psi$	$\neg\psi$	$\neg(\varphi \leftrightarrow \psi)$	$\neg(\varphi \wedge \psi)$
T	T	F	F	F	F	F	F	F	F
T	F	F	F	F	F	T	T	T	T
F	T	F	F	T	T	F	F	T	T
F	F	F	T	F	T	F	T	F	T

φ	ψ	$\varphi \wedge \psi$	$\varphi \leftrightarrow \psi$	$\psi \rightarrow \varphi$	φ	$\varphi \rightarrow \psi$	$\varphi \vee \psi$	$\varphi \vee \neg\varphi$
T	T	T	T	T	T	T	T	T
T	F	F	F	F	T	F	T	T
F	T	F	F	T	F	T	T	T
F	F	F	T	F	F	T	F	T

But the number of truth-functions grows doubly exponentially with the number of variables, so there are 256 truth-functions of three variables, 65,536 truth-

functions of four variables, *more than four billion* truth-functions of five variables, and so on. Now our meager collection of five connectives starts to look inadequate for the task. But, in fact, it is more than adequate to express *any possible* truth-function of *any number* of variables. The trick is that *any formula whatsoever* of **PL** can be written in **disjunctive normal form**, which uses only disjunction, conjunction, and negation.

You can construct the disjunctive normal form for an arbitrary formula by following this procedure:

1. Construct the truth table for the formula.
2. For every row where the column under the main connective for the formula has a **T**, build a conjunction of each of the atomic sentences in the formula, negating each atomic sentence that has an **F** in that row (and leaving the ones that have a **T** unnegated).
3. Disjoin all those conjunctions.
4. *Special case*: if there is no row with a **T** under the main connective (that is, if the formula is a contradiction), the procedure yields no conjunctions to disjoin. In that case, take any atomic sentence letter A occurring in the formula and use $A \wedge \neg A$ as its disjunctive normal form.

Let us build the disjunctive normal form for the conditional $A \rightarrow C$. We know that the truth-table for this formula will have **T**s under the conditional in all rows except the one where A is true and C is false:

A	C	$A \rightarrow C$		
T	T	T	T	T
T	F	T	F	F
F	T	F	T	T
F	F	F	T	F

[Try it in the Logic Lab ↗](#)

Now that we have the truth-table, we form all the relevant conjunctions: $A \wedge C$, $\neg A \wedge C$, and $\neg A \wedge \neg C$. Finally we disjoin them: $(A \wedge C) \vee (\neg A \wedge C) \vee (\neg A \wedge \neg C)$. This last formula is guaranteed to have the same truth-table as $A \rightarrow C$. In general, by construction the disjunctive normal form of a formula is guaranteed to have the same truth-table as the formula we began with. (If the formula has just one

row with a **T**, the “disjunction” consists of a single conjunction; and if it has none, the special case above applies.) Because we can build the disjunctive normal form for *any* formula, we are guaranteed that we can express any formula using just negation, conjunction, and disjunction.

This means, of course, that we can define conditional and biconditional in terms of disjunction, negation, and conjunction. A brute-force way of defining them is by constructing their disjunctive normal forms (we just did that for the conditional). But there are alternative definitions that are more perspicuous. For instance, we can define the conditional $\varphi \rightarrow \psi$ as $\neg\varphi \vee \psi$ and also as $\neg(\varphi \wedge \neg\psi)$. Moreover, we don’t really need all three of negation, conjunction, and disjunction, for disjunction can be defined in terms of negation and conjunction, and conjunction can be defined in terms of negation and disjunction.⁵ Thus, with just two of our five connectives we can express any formula of **PL**.

We can do even better in terms of expressive economy, for we can express any formula of **PL** with just *one* connective—although not any of our list of five. One connective which is functionally complete is the *Sheffer stroke*, or NAND, which is usually symbolized with " $\uparrow$ " and has the following truth-table:

Table 1.8: Truth table for the Sheffer stroke

φ	ψ	$\varphi \uparrow \psi$
T	T	F
T	F	T
F	T	T
F	F	T

It is easy to see from the truth-table why the Sheffer stroke is also called NAND, because $\varphi \uparrow \psi$ has the same truth-table as $\neg(\varphi \wedge \psi)$. With NAND, we can define negation, conjunction, and disjunction,⁶ so it follows from what we know about disjunctive normal form that we can define every possible formula of **PL** with NAND alone.

There is another binary connective which is also functionally complete: NOR, usually symbolized with " $\downarrow$ " and with the following truth-table:

⁵See the exercises below.

⁶See exercises below.

Table 1.9: Truth table for NOR

φ	ψ	$\varphi \downarrow \psi$
T	T	F
T	F	F
F	T	F
F	F	T

NOR is so-named because $\varphi \downarrow \psi$ has the same truth-table as $\neg(\varphi \vee \psi)$. As with NAND, so too with NOR: with it, we can define negation, conjunction, and disjunction⁷—and, therefore, it too is functionally complete.

Given that a single connective is functionally complete, why do we bother to introduce five? Because there is a trade-off between expressive power and legibility (for us humans). Consider how a relatively simple formula like $A \rightarrow (B \vee C)$ must be expressed using just NAND: $A \uparrow ([(B \uparrow B) \uparrow (C \uparrow C)] \uparrow [(B \uparrow B) \uparrow (C \uparrow C)])$. This formula is extremely difficult to parse, whereas the original one is relatively trivial. Using just one connective gives us a lot of expressive power in a small package, but it does so to the detriment of legibility. That’s why we use five connectives: because we are interested in *using* **PL**. Why would anyone use only one connective, then? Because using just one connective makes it much easier to prove things *about* **PL**—something that logicians are very interested in.

Exercises

1. Give an example of three different sentences that express the same proposition.
2. Don’t we know that some sentences in the interrogative mood express propositions?
3. Why is it that for n atomic propositions there are 2^n combinations of truth-values?
4. What is wrong with defining conjunction as follows?

⁷Again: see exercises below

A	B	$A \wedge B$
T	T	T
T	F	F
F	T	F
F	F	F

5. Define conjunction in terms of negation and disjunction, and disjunction in terms of negation and conjunction.
6. Define negation, conjunction, and disjunction in terms of NAND. Do the same in terms of NOR.
7. For each of the following formulas: determine whether it is a wff according to our strict definition; determine whether it is a wff when considering the conventions for dropping parentheses; for each wff (strict or not), construct its syntax tree and its truth-table and determine whether it is a tautology, a contradiction, or a contingency.

- a) $((\neg B) \vee (A \rightarrow C))$ Practice in the Logic Lab ↗
- b) $\neg A \leftrightarrow ((\neg A \wedge B) \vee (\neg A \wedge \neg B))$ Practice in the Logic Lab ↗
- c) $A \neg(B \vee C)$ Practice in the Logic Lab ↗
- d) $\neg \neg A \vee (B \wedge A)$ Practice in the Logic Lab ↗
- e) $(A \leftrightarrow B \vee C)$ Practice in the Logic Lab ↗
- f) $(A \rightarrow B) \vee (B \rightarrow A)$ Practice in the Logic Lab ↗
- g) $(A \wedge \neg A) \leftrightarrow (B \wedge \neg B)$ Practice in the Logic Lab ↗
- h) $(A \vee B) \rightarrow (A \wedge B)$ Practice in the Logic Lab ↗
- i) $(\neg A \wedge (B \vee C)) \rightarrow (C \vee A)$ Practice in the Logic Lab ↗
- j) $\neg(A \vee B) \rightarrow C \wedge$ Practice in the Logic Lab ↗

2 Translations

*This chapter explains how to translate from English to **PL**.*

2.1 From English to PL

Recall that we constructed the artificial language of **PL** in order to give a toy model of validity for arguments. In order to do that, we need to learn to translate between English and **PL**.

The following table gives you a preliminary translation guide for our five connectives of **PL**:

Table 2.1: Truth-functional connectives and common English counterparts

PL connective	Typical English counterpart
$\neg$	not
$\wedge$	and
$\vee$	or (inclusive)
$\rightarrow$	if ... then ...
$\leftrightarrow$	if and only if

In order to translate a sentence from English to **PL** we need to first identify whether it is an atomic or a molecular sentence. If it is an atomic sentence, then our job is easy: we just pick an atomic sentence letter to stand for it. If it is a molecular sentence, then our job is more complex: first we need to identify which atomic sentences the molecular sentence is made up of (and assign them corresponding sentence letters), next we have to identify which connectives the sentence uses and which sub-sentences those connectives connect, and finally we connect the sub-sentences of **PL** to each other with the corresponding connectives.

Let us start with the easy case:

Leona is a cat.

That sentence is atomic: it doesn't have any other sentences as parts. Therefore, our translation will be pretty easy:

Dictionary:

P = Leona is a cat.

But consider now this rather more complicated sentence:

Leona is not a cat or she is a Maine Coon if and only if Lucas likes to play with her if she is purring.

The first thing we need to do is to identify the atomic sentences that make up that molecular sentence. To do that it will actually help to start with the second step and identify the connectives we see. In this case, we have "not", "or", "if and only if" and "if". Notice that *Leona is not a cat* is not an atomic sentence, but a molecular one. Negation is the only unary connective, which forms a molecular sentence out of a single sentence. We have therefore identified our first atomic sentence, which is the same as before, "Leona is a cat". We also have "She is a Maine Coon", "Lucas likes to play with her" and "She is purring". Those sentences have pronouns ("she", "her"), and in isolation it will not be clear who those pronouns are referring to. Therefore, when building our dictionary we should replace those pronouns with the name of the person or animal they are referring to (in this case, "Leona"). So now we have our dictionary:

Dictionary:

C = Leona is a cat.

M = Leona is a Maine Coon.

L = Lucas likes to play with Leona.

P = Leona is purring.

Next we need to figure out how those atomic sentences are connected to each other via our connectives. Clearly, the negation is affecting C , so that we should somewhere have $\neg C$ as the translation of "Leona is not a cat". Next, we need to figure out which one is the main connective of the whole sentence. It cannot be the negation, because that has scope over just one atomic sentence, not the whole molecular one. The other three connectives all have some claim to being the main one. For instance, we could perhaps take the disjunction to be the main connective of the sentence. But if whoever wrote (or uttered) this sentence wanted disjunction

to be the main connective, he could have done a much better job of conveying his intentions by writing instead:

Either Leona is not a cat, or else she is a Maine Coon if and only if Lucas likes to play with her if she is purring.

Given the sentence as we have it, the best interpretation is that disjunction is not the main connective. We are left with two options: the biconditional or the conditional. The conditional here is indicated by an "if" unaccompanied by a "then". How are we to interpret it? As we will see soon, this happens often ("If it's raining, the picnic is cancelled"). In general, what follows the "if" is the antecedent of the conditional (unless we have an "only if", in which case it is the consequent). So P will be the antecedent of the conditional. What is its consequent? If the conditional were the main connective, the consequent will have to be "Leona is not a cat or she is a Maine Coon if and only if Lucas likes to play with her". That would be a very strange information to want to transmit. Let's try the biconditional as the main connective. The right-hand side of this biconditional is "Lucas likes to play with [Leona] if she is purring", which we will translate as $P \rightarrow L$. And what would be its left-hand side? "Leona is not a cat or she is a Maine Coon", of course. Putting it all together, we have:

$$(\neg C \vee M) \leftrightarrow (P \rightarrow L)$$

This exercise has brought to the fore a very important lesson: natural language sentences are often ambiguous. We want to interpret the sentence so as to reflect the intentions of the speaker or writer. Thankfully, most of the time the ambiguities can be resolved by paying attention to the context in which the sentence is used. Sometimes, however, there will be unresolvable ambiguities. In that case, either the user of the sentence simply did a bad job, or (more often) we are just missing important context that helps with disambiguation.

2.2 Translation notes for connectives

We have above a table with basic information about translating from English to **PL**. However, natural languages all have different ways of transmitting information, and this applies to our five connectives. Let's look at some of those.

2.2.1 Negation

As we saw before, negation in English can be expressed by a simple "not", as in "Leona is not a cat". Remember that contractions, such as "didn't" and "can't", also contain a "not". There are also more bombastic expressions of negations, such as "it's not the case that", or "it is false that".

Table 2.2: Common English expressions of negation

English expression	PL translation
not φ (and contractions)	$\neg\varphi$
it is not the case that φ	$\neg\varphi$
it is false that φ	$\neg\varphi$

2.2.2 Conjunction and its English cousins

The most common English expression translated into **PL** as " $\wedge$ " is, of course, "and". But there are important differences between "and" and " $\wedge$ " that we should be aware of. To begin with, "and" is often used to express temporal order in addition to conjunction. For instance, when we say "John walked in and took a seat" what we mean can be more explicitly expressed by saying "First John walked in, and then he took a seat." The " $\wedge$ " of **PL** cannot express this temporal order, for an interesting reason. The reason is that temporal order is not truth-functional. Suppose that it is true both that John came in and that he took a seat. It may in that case be true that he came in *first* and took a seat *afterwards*, but it may also be false. Even setting aside the ambiguity of "taking a seat", which can mean picking up a chair in addition to sitting on it, it may be that John first sat and then came in (hopping while seated in his chair, perhaps). So, the truth values of "John came in" and "John took a seat" do not settle the truth value of "First John came in and afterwards he took a seat" or of "John came in and took a seat" in its obvious interpretation. That means that "and" in English sometimes has a non-truth-functional interpretation, and we know that **PL** only deals in truth-functional connectives.

Besides "and", other English constructions can be translated into **PL** by " $\wedge$ ". One of them is "but". But "but" carries with it some additional information that is not part of its truth conditions, and that therefore is not captured by " $\wedge$ ". Consider, for instance, the sentence "She loves him, but she won't marry him." The "but" is not just conjoining "She loves him" with "She won't marry him", but rather also introducing a clashing contrast between them (suggesting that there is something

surprising about the fact that she won't marry him given that she loves him). Nevertheless, "She loves him, but she won't marry him" is true just in case she loves him and she won't marry him, so as far as truth conditions go "but" behaves just like "and". The contrast is an additional element that the word conveys by convention (philosophers call this a *conventional implicature*), not a contribution to the sentence's truth conditions. Another English conjunction that works in a similar way is "yet" ("he is thin yet strong"). Some textbooks will tell you that the following should also be translated by " $\wedge$ ": "moreover", "however", and "nonetheless". However, these English expressions often function as *discourse connectives* rather than sentence connectives. That is to say, they are often used to introduce a new sentence, one that contrasts in some way with the previous one, rather than to add a conjunct. This nevertheless highlights an interesting fact: there is little difference between separately asserting two sentences and asserting their conjunction (at least as far as truth conditions go).

Table 2.3: Common English expressions of conjunction

English expression	PL translation
φ and ψ	$\varphi \wedge \psi$
φ but ψ	$\varphi \wedge \psi$
φ yet ψ	$\varphi \wedge \psi$

2.2.3 Disjunction—inclusive and exclusive

The " $\vee$ " of **PL** has a clear English companion in "or" (sometimes accompanied by "either"). Recall, however, that we said that " $\vee$ " is an *inclusive* disjunction. What that means is that a disjunction is true whenever at least one of its disjuncts is true, including when both are. With the English "or" this is sometimes not the case. Suppose that the waiter tells you: "You can have soup or salad with your entree". Usually, that means that you cannot have both (for the same price). That use of the English "or" is an exclusive one, meaning that the sentence is false when both of its disjuncts are true. Thus, it would be wrong to translate that sentence as " $P \vee Q$ ". However, the exclusive "or" of English *can* be translated into **PL**—just not with a simple " $\vee$ ".¹

¹How? See the exercises.

2.2.4 Conditional

The conditional connective of **PL**, " $\rightarrow$ ", has given rise to notorious philosophical disputes that center on whether it is a faithful rendition of the English indicative conditional, often expressed with an "if ... then ..." construction. Recall from our truth table for the conditional that it is false exactly in one case: when its antecedent (the formula to the left of the arrow) is true and its consequent (the sentence letter to the right of the arrow) false. This means, in particular, that a conditional of **PL** is automatically true whenever its antecedent is false or its consequent true. Does that happen also with the English conditional? For instance, in a context where it is assumed that it is not raining, is it true that if it is raining then Mbappé is a better soccer player than Messi? It may seem that the answer here is a clear "No", regardless of what you think of the truth-value of the consequent. But things are complicated.

One clear reason why no one would utter the sentence "If it is raining then Mbappé is a better soccer player than Messi" is that English conditional constructions usually express some kind of connection between antecedent and consequent, and there is no discernible connection in this case. Notice that if this is the correct explanation for why we find that sentence odd, it has little to do with whether the antecedent is true or false. What does rain have to do with the comparison between Mbappé and Messi?² Notice that conditionals where a connection is plausible do not sound bad. Suppose we have just woken up, and we do not know whether it is raining or not. I say "If it is raining, then the streets are wet". We then look outside the window and notice that it is not raining. My previous assertion seems still to be perfectly correct. Still, this implied connection between antecedent and consequent is not something that can be captured in **PL**, because it is another example of an aspect of the meaning of the English conditional that goes beyond the truth conditions captured by the truth-table of $\rightarrow$ (on many views, it is not part of the truth conditions at all).

Besides the "If ... then ..." construction, conditionals can be expressed in other ways in English. For instance, I can say "The streets are wet, if it is raining". This expresses the same proposition as "If it is raining, then the streets are wet", just with the antecedent coming after the consequent. However, if I say "The streets are wet only if it is raining" I express a different proposition, one where the antecedent

²In a context where it is assumed that it is indeed raining—as opposed to a context where it is assumed that it is not or a context where neither of those two things is assumed—asserting the conditional can be seen as a merely embellished way of asserting its consequent.

and consequent have changed places. "If it is raining, then the streets are wet" is false exactly when it is raining and the streets are not wet, whereas "The streets are wet only if it is raining" is false exactly when the streets are wet and it is not raining. That second sentence, for instance, is false when it is not raining but the streets are wet because they are being cleaned, for example, whereas the first sentence would not be false in that case.

English contains conditionals other than the indicative one. Consider, for instance, the following one: "If Argentina had won the World Cup in 2026, Juan would have been unhappy". This is a subjunctive conditional, and clearly does not behave like the indicative conditional—nor like the conditional of **PL**. It is generally accepted that the subjunctive conditionals like that one are not truth-functional connectives, and so cannot be translated into **PL**.

Table 2.4: Common English expressions of the material conditional

English expression	PL translation
if φ , then ψ	$\varphi \rightarrow \psi$
φ if ψ	$\psi \rightarrow \varphi$
φ only if ψ	$\varphi \rightarrow \psi$
φ is a sufficient condition for ψ	$\varphi \rightarrow \psi$
φ is a necessary condition for ψ	$\psi \rightarrow \varphi$

2.2.5 Biconditional

The biconditional of **PL** is simply a conditional going in both directions. If you look at its truth-table, what a biconditional says is that both of its sides share a truth-value: either they are both true or both false. In other words, a biconditional is false exactly when one of its sides is true and the other false.

As far as its expression in English goes, "... if and only if ..." is the canonical way of stating a biconditional. Other expressions are "... just in case ..." and "... exactly when ..."—although it is fair to say that one rarely encounters biconditionals in natural languages outside of mathematical or logical contexts.

Table 2.5: Common English expressions of the material biconditional

English expression	PL translation
φ if and only if ψ	$\varphi \leftrightarrow \psi$
φ just in case ψ	$\varphi \leftrightarrow \psi$
φ exactly when ψ	$\varphi \leftrightarrow \psi$
φ is a necessary and sufficient condition for ψ	$\varphi \leftrightarrow \psi$

2.2.6 Some other expressions

A few other English expressions come up often enough that they deserve a mention. In each case, the translation below is the standard one, but as always you should check that it fits the context in which the sentence is used.

Unless. “ φ unless ψ ” says that if ψ does not happen, then φ does. So it is translated as $\neg\psi \rightarrow \varphi$, which has the same truth-table as $\varphi \vee \psi$. For instance, “Bob will not like the strawberries unless they are ripe” can be translated as $\neg R \rightarrow \neg L$ or, equivalently, as $\neg L \vee R$. Notice that this treats “unless” as an *inclusive* disjunction: nothing is said about what happens when ψ is true.³

Neither . . . nor. “Neither φ nor ψ ” says that φ is false and ψ is false, so it is translated as $\neg\varphi \wedge \neg\psi$, or equivalently as $\neg(\varphi \vee \psi)$. Do not confuse it with “not both φ and ψ ”, which is translated as $\neg(\varphi \wedge \psi)$ and is compatible with one of them being true.

Provided that, as long as, on the condition that. These all introduce the antecedent of a conditional, just like “if”: “ φ provided that ψ ” is translated as $\psi \rightarrow \varphi$.

The only way. “The only way for φ to happen is for ψ to happen” says that φ cannot happen without ψ , so it is translated as $\varphi \rightarrow \psi$, exactly like “ φ only if ψ ”.

³Sometimes “unless” is meant exclusively, as in “You will get dessert unless you refuse to eat your vegetables”, which also suggests that if you refuse you won’t get dessert. As with “or”, this extra information can be captured by using a biconditional or the exclusive disjunction from the exercises.

Table 2.6: Some other English expressions and their **PL** translations

English expression	PL translation
φ unless ψ	$\neg\psi \rightarrow \varphi$ (equivalently, $\varphi \vee \psi$)
neither φ nor ψ	$\neg\varphi \wedge \neg\psi$ (equivalently, $\neg(\varphi \vee \psi)$)
not both φ and ψ	$\neg(\varphi \wedge \psi)$
φ provided that ψ	$\psi \rightarrow \varphi$
the only way for φ is ψ	$\varphi \rightarrow \psi$

Try it in the Logic Lab ↗

Exercises

1. How would you define exclusive disjunction?

Translate the following English sentences into PL. Define any atomic sentences that you are going to use. [Practice translating in the Logic Lab](#) ↗

2. Alyssa will play basketball if the weather is nice and she doesn't have work.
3. Bob will not like the strawberries unless they are ripe.
4. Charly graduates if and only if she's satisfied her curriculum requirements and she has enough credits.
5. Deion will fall if the stairs are slippery or wet.
6. Erin likes to watch TV shows and movies but does not like to play basketball.
7. Francis won't go to the party just in case neither his mother nor his father gives their permission.
8. Gianna goes hiking if the weather is good, but, if the weather isn't good, she plays video games.
9. The only way Hadrian will get a trophy is if either his soccer team or tennis team win their championships.
10. Ida will be successful if she works hard or if she is lucky.
11. Joel gets sick if he eats ice cream or tomatoes without taking his medication.

3 Arguments

This chapter examines arguments and the notions of validity and soundness. It also explains how to use truth-tables to determine whether an argument is tautologically valid.

3.1 Good Arguments

An argument, in the sense used in philosophy, is an attempt to rationally convince an audience of the truth of a proposition (the *conclusion* of the argument) on the basis of some other propositions (the *premises* of the argument). Some comments about this characterization are in order. First, the audience of the argument might be one person (for instance, your friend, when you are trying to convince him of something) or several people (for instance, when a politician is trying to convince an electorate to vote for him). It can also happen that the audience of an argument is the very same person who constructs the argument. Sometimes, we “argue with ourselves” in order to figure out what to believe. Second, notice that the kind of conviction that an argument aims to impart in its audience is *rational* conviction. This is why many times appeals to emotions and other rhetorical devices are classified as *fallacies* (i.e., attempts to produce conviction through non-rational means).

What, therefore, is a *good* argument? A good argument is one which succeeds at what it aims: i.e., a good argument is one which gives its audience good reason to believe its conclusion (notice that we characterized an argument as an *attempt* to produce rational conviction, and attempts can fail). Whether an argument is good does not depend on whether anyone actually ends up convinced by it: people can fail to be convinced by good arguments, and can be convinced by bad ones. What matters is whether the audience *should* be convinced. For that to be so, two things must hold: first, the audience must be rationally justified in accepting its premises, and second, the premises must offer good reasons to accept the conclusion. Notice

that, because of the first condition, goodness is relative to an audience: the same argument can be good for an audience that is justified in accepting its premises and not good for one that is not.

Recall the two arguments we started out with:

- (1) Leona is a cat.
- (2) If Leona is a cat, then Leona is an animal.

∴ Leona is an animal.

- (1) Leona is an animal.
- (2) If Leona is a cat, then Leona is an animal.

∴ Leona is a cat.

If you are justified in believing the premises of the first argument (i.e., that Leona is a cat, and that if that is true, then she is an animal), that gives you rational justification for believing its conclusion. So that first argument is well-positioned to be a good one. However, even if you are justified in believing the premises of the second argument (i.e., that Leona is an animal and that if Leona is a cat then she is an animal), that does not give you good reasons for believing its conclusion (that Leona is a cat). There are so many different kinds of animals that it would not be rational to believe that Leona is a cat solely on the basis that she is an animal and that cats are animals. So that second argument is not a good one.

3.2 Validity

Notice an important difference between the first and the second argument: in the first argument, it is impossible for its premises to be true and its conclusion false; in the second argument, it is possible for the premises to be true and the conclusion false (for instance, both premises will be true if Leona is a dog, but its conclusion will be false). We will call this property that the first argument has and the second lacks, *validity*. We must now make an important clarification regarding the notion of validity.

Validity is defined in terms of *possibility*, but there are different notions of possibility. In one sense, it is impossible for us to transport a million humans to Mars this year—but it is impossible only in the sense that we cannot do it with the

technology available to us today. Doing so does not violate the laws of physics. So we say that transporting a million humans to Mars is technologically impossible, but physically (or nomologically) possible.¹ Faster than light travel, on the other hand, is nomologically impossible, not just technologically impossible. But it is not logically impossible—the proposition *This ship moves faster than the speed of light* is not a contradiction. So, as far as possibilities are concerned, everything that is technologically possible is nomologically possible, and everything that is nomologically possible is logically possible:

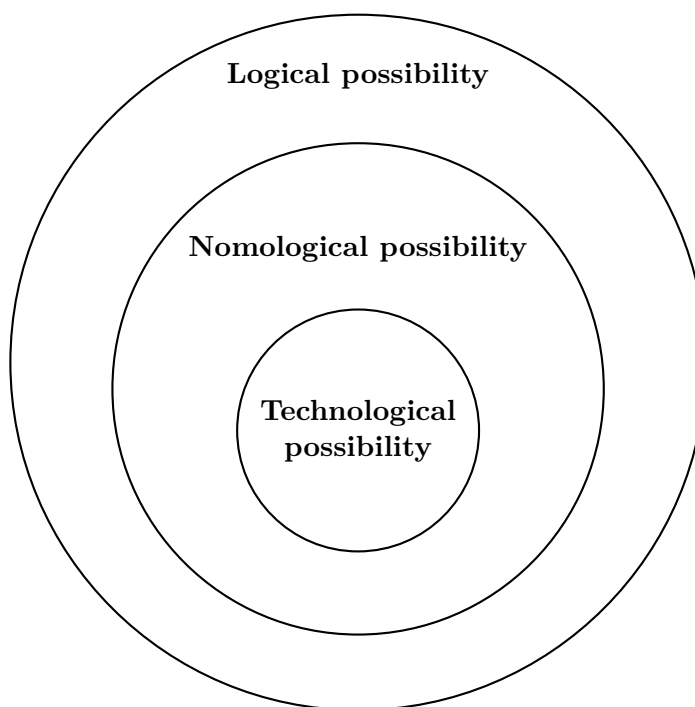


Figure 3.1: Different notions of possibility

Impossibility, of course, works the other way around: whatever is logically impossible is nomologically impossible, and everything that is nomologically impossible is technologically impossible:

¹Notice that technological possibility is implicitly relative to a time: airplanes were once technologically impossible, and transporting humans to Mars will one day be technologically possible.

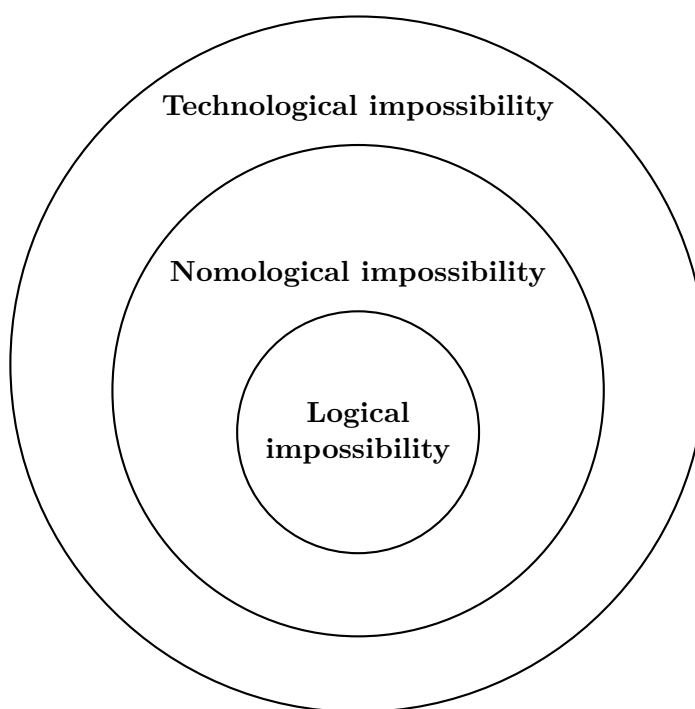


Figure 3.2: Different notions of impossibility

Philosophers sometimes talk of yet a fourth notion of possibility, *metaphysical* possibility, but we will not be concerned with that one here. So, when we define validity in terms of possibility, which notion of possibility are we using? Logical possibility. When we say that an argument is valid, we mean to say that the proposition asserting all of its premises but negating its conclusion is a contradiction.

Notice that when we say that an argument is valid we are not saying anything specific about the truth-values of its premises and conclusion. What we are saying, again, is that it is impossible for a valid argument to have all true premises and a false conclusion. So, if we are told that the premises of an argument are in fact all true and its conclusion false, that guarantees that the argument is invalid. But any other combination of truth-values of premises and conclusion does not determine whether the argument in question is valid or invalid. For instance, there are arguments with true premises and a true conclusion which are invalid: even though their premises and conclusion are *in fact* true, it is possible for its conclusion to be false while its premises remain true. Consider, for instance, the second argument we have been talking about:

- (1) Leona is an animal.

(2) If Leona is a cat, then Leona is an animal.

$\therefore$ Leona is a cat.

As it happens, this argument has true premises and a true conclusion. But it is *possible* for its premises to be true and its conclusion false. How can we tell that it is possible? Well, a good test is the following: can we imagine a situation where the premises are true and the conclusion false, without falling into a contradiction? Yes we can: we just have to assume that Leona is a dog. Then both premises are true, and the conclusion false, and there is no contradiction. By contrast, there is no way to imagine a situation where Leona is a cat and if Leona is a cat, then Leona is an animal but Leona is not an animal without falling into a contradiction.

Similarly, an argument with false premises and a true or false conclusion can be either valid or invalid, depending on whether it's *possible* for its premises to be true and its conclusion false. We can summarize the relationship between validity and invalidity of arguments and truth and falsity of premises and conclusion in the following table:

Table 3.1: Validity and possible truth-value combinations

	Premises true		Premises false	
	Conclusion true	Conclusion false	Conclusion true	Conclusion false
Valid	✓	×	✓	✓
Invalid	✓	✓	✓	✓

The definition of validity has two consequences that students often find surprising. Consider first this argument:

(1) Leona is a cat.

(2) Leona is not a cat.

$\therefore$ Lucas plays the trumpet.

Is it possible for the premises to be true and the conclusion false? No, because it is not possible for both premises to be true at all: no logically possible situation is one where Leona is a cat and is not a cat. So there is no situation where the premises are true and the conclusion false, and the argument is valid. In general, an argument whose premises cannot all be true together is valid, whatever its conclusion is. We call such arguments *vacuously valid*.

Now consider this argument:

(1) Lucas plays the trumpet.

∴ Either Leona is a cat or Leona is not a cat.

Here the conclusion cannot be false, so again there is no possible situation where the premises are true and the conclusion false, and the argument is valid. In general, an argument whose conclusion cannot be false (for instance, a tautology) is valid, whatever its premises are.

These arguments are valid, but they are not good. In the first one, the premises cannot both be true, so it is not an argument that could rationally convince anyone of anything, and in the second one the premise does no work at all. This is one more illustration of the point of the next section: validity is not enough for an argument to be good.

3.2.1 Validity, soundness and good arguments

We introduced the notion of validity, and we will soon see how we can use truth-tables to check arguments for *tautological* validity (i.e., the kind of validity associated with **PL**). But we need to pause to make a very important point: not all valid arguments are good, and not all good arguments are valid.

It is easy to see that not all valid arguments are good, for some valid arguments have obviously false premises. Given that a good argument must rationally convince its audience of its conclusion on the basis of its premises, an argument with obviously false premises cannot do that. But what about valid arguments with true premises? We call these **sound** arguments. Aren't all sound arguments good? Not really. Suppose that you and a friend are trying to convince each other of your respective beliefs: you believe that God exists, your friend believes that God doesn't exist. One of these arguments is guaranteed to be sound, and yet neither one of them should rationally convince either you or your friend of their respective conclusions:

(1) God exists.

∴ God exists.

(1) God does not exist.

∴ God does not exist.

We have established, then, that not all valid arguments (and even not all sound arguments) are good. What about the converse, though? Even though being valid is not a sufficient condition for being a good argument, couldn't it be a necessary condition? That is to say: must every good argument be valid? No. There are good invalid arguments, for there are arguments where, while it is *possible* for their premises to be true and their conclusion false, this is highly unlikely. If this is recognized by the audience, then the argument may well convince its audience of believing its conclusion on the basis of its premises despite its invalidity. This is what happens, for instance, with the evidence we have for our best scientific theories: even though the argument from the empirical data to those theories is not valid, belief in those theories (or at least the best established parts of them) is still one of the best supported beliefs we have.

But we do not have to wander into the realm of scientific theories to find good invalid arguments. Suppose you believe that Leona is an animal because I tell you that Leona is a cat and if Leona is a cat then she is an animal. That argument is valid, but your degree of justification in its conclusion cannot surpass the degree of justification you have for believing in its premises. And what justifies you in believing its premises? Well, what justifies you in believing that Leona is a cat is that I told you so (and maybe even showed you some pictures of her). But why do you believe me? Is there a good valid argument whose conclusion is that I never lie? Even I find that hard to believe. Your justification for believing that Leona is a cat, therefore, does not come from a valid argument. And yet you are surely justified in believing that (trust me!). Some will think that what justifies you is not an argument, but many will think that it is an invalid argument—perhaps along the following lines:

- (1) Juan says that Leona is a cat.
- (2) I have no specific reason to distrust Juan on this occasion.
- (3) If I have no specific reason to distrust someone when they say something, then it's likely that what they say is true.

∴ Leona is a cat.

If not all good arguments are valid and not all valid arguments are good, why are we interested in this notion of validity? One reason to be interested in it is that validity can be seen as one end of the spectrum in what kind of support the

premises of an argument provide to its conclusion. To be good, an argument's premises must provide good reasons for thinking that its conclusion is true. If it is impossible for its premises to be true and its conclusion false (i.e., if the argument is valid), this is the strongest kind of reason the premises can give. But the goodness of an argument depends on more than what support the premises provide to the conclusion: it depends as well on how justified we are in believing its premises. So, we study validity because it is an important end-point in one of the factors that make for a good argument. In the second part of this class we study probability, which many see as key in understanding the less-than-perfect support that some premises provide to an argument's conclusion.

We can summarize the relationships between validity, soundness, and good arguments as follows:

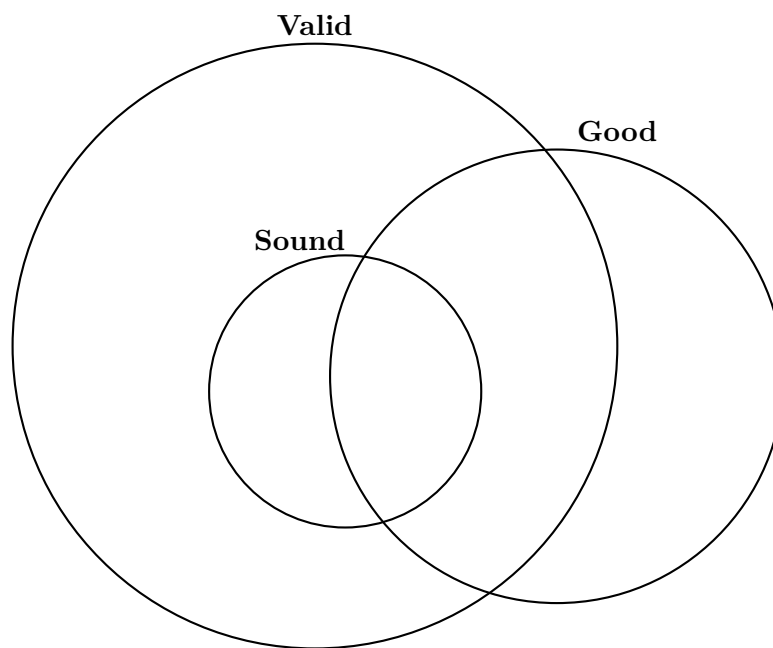


Figure 3.3: Valid, sound, and good arguments

3.3 Tautological Validity

Just as we use the language of **PL** to represent some of the logical structure of natural languages such as English, we will use truth-tables to model the notion of validity. Recall that we say that an argument is valid just in case it is impossible for its premises to be true and its conclusion false. In other words, every possibility

where all its premises are true is a possibility where its conclusion is true as well—or, equivalently, there is no possibility where all its premises are true and its conclusion false.

What are these possibilities we are talking about? Philosophers like to talk about *possible worlds*—consistent and complete ways things could have been. Suppose we think there could have been a talking donkey—i.e., we think that a talking donkey is a possibility. What that means is that there is a possible world that includes a talking donkey—i.e., there is a consistent and complete way things could have been that include a talking donkey. When we say that a way things could have been is *consistent* we mean that no contradiction follows from a description of that way things could have been. And when we say that a way things could have been is *complete* we mean that it answers every yes/no question we can ask about it.

PL gives us a model of possible worlds. Suppose we are working with n atomic propositions. As we know, that means that there are 2^n possible combinations of truth-values. For instance, if we are working with just these two atomic propositions, *Leona is a cat* and *Lucas plays the trumpet*, then they can be both true, both false, or one of them true and the other false, with four total possible combinations of truth-values. In **PL**, each one of these possible combinations of truth-values represents a possible world. That is to say, in a model with just two atomic propositions there are only four possible worlds—four ways to answer consistently every yes/no question we can ask about every possible truth-functional combination of those two atomic propositions. Of course, the number of possible worlds grows exponentially with the number of atomic propositions.

When we are evaluating an argument, we need to deal only with those possible worlds determined by the atomic propositions involved in the argument—we can safely ignore every other possible world. So, in order to check whether an argument is *tautologically valid* we follow this procedure: first, we construct a *combined truth-table* for all the premises and the conclusion of the argument; second we check whether there is a row of this combined truth-table where all the premises are true but the conclusion false; if there is such a row, the argument is tautologically invalid, if there isn't, the argument is tautologically valid. Let's illustrate with our two working examples. We start by translating the first argument to **PL**:

Dictionary

C = Leona is a cat.

A = Leona is an animal.

Formalized argument

- (1) C
 (2) $C \rightarrow A$
 $\therefore A$

Then we make the combined truth-table:

C	A	Premise 1 C	Premise 2 $C \rightarrow A$	Conclusion A
T	T	T	T	T
T	F	T	F	F
F	T	F	T	T
F	F	F	T	F

[Try it in the Logic Lab ↗](#)

Notice that the only row of the combined truth-table where both premises are true is row 1, and in that row the conclusion is true as well. Therefore, the argument in question is tautologically valid.

Now let us do the same for the second argument:

Dictionary

C = Leona is a cat.

A = Leona is an animal.

Formalized argument

- (1) A
 (2) $C \rightarrow A$
 $\therefore C$

C	A	Premise 1 A	Premise 2 $C \rightarrow A$	Conclusion C
T	T	T	T	T
T	F	F	F	T
F	T	T	T	F
F	F	F	T	F

Try it in the Logic Lab ↗

In this case, there is a row where all the premises are true and the conclusion false: row 3. Therefore, the argument is invalid. That third row is a *counterexample* for that argument.

3.4 Translating arguments

The first step in determining whether an argument is tautologically valid consists in translating it from English to **PL**. By now we know how to translate single propositions into **PL**, but how do we translate an argument? Well, an argument is built up of propositions, so the first step is just to translate every proposition in the argument. The second, more difficult step, is to determine which of those propositions is the conclusion (the others, of course, will be the premises).² When we encounter arguments in the wild (as opposed to in a textbook), they do not come neatly arranged with the premises first and the conclusion clearly presented as such. The conclusion might come first, or in the middle, the premises may be mixed in with argumentatively irrelevant data, some premises may be merely implicit, etc. We can be helped in identifying the premises and conclusion of an argument by certain phrases that commonly introduce premises, and others that commonly introduce the conclusion (see the table below). But identification of arguments is part of reading (and listening) comprehension, and can be improved with practice.

Table 3.2: Common premise and conclusion indicators

Premise indicators	because, since, for, given that, the reason is, after all
Conclusion indicators	therefore, so, thus, hence, consequently, it follows that, must

Some of these words are better indicators than others. In particular, “must” is a weak indicator of a conclusion: it often marks a conclusion that the speaker has inferred (“Luke must have not paid attention in class”), but it can also express necessity or obligation without signaling any argument at all (“You must be home by ten”). Use context to decide.

²Not every piece of discourse will constitute an argument, so determining whether we are faced with an argument or some other kind of discourse is the zeroth step.

Let us end this part with a reasonably complex argument in English, and let us translate it into **PL** and then use truth-tables to find out whether it is tautologically valid. The following is a reconstruction of the reasoning in a Supreme Court decision in 1952.³ That year, President Truman issued an order directing the Secretary of Commerce to seize the nation's steel mills during the Korean War to prevent a strike. The steel companies sued the Secretary of Commerce, Charles Sawyer, arguing that the order was an unconstitutional exercise of lawmaking power by the President. The case reached the Supreme Court, which agreed with the companies in an opinion written by Justice Hugo Black. (Justice Robert Jackson wrote a famous concurring opinion with a different analysis, which we will not discuss.) In outline, the reasoning of the majority was the following; what follows is a summary of the argument, not a quotation from the opinion:

The President's authority to seize the steel mills must come either from Congress or from the Constitution. No statute expressly authorizes the seizure, nor can such authority fairly be implied from any statute. The seizure cannot be justified by the President's power as Commander in Chief. Nor can it be justified by the other executive powers granted to the President by Article II.

The first step in translating this argument into **PL** is to figure out what the conclusion is. In this case, the conclusion is not explicitly presented in the passage, but it is clear that it must be something like "The President did not have the legal authority to seize the steel mills." Once the conclusion is identified, it is easy to see what the premises are. We can now put the argument in a more perspicuous form:

- (1) If the President had the legal authority to seize the steel mills, then that authority came either from Congress or from the Constitution.
 - (2) The authority did not come from Congress.
 - (3) The authority did not come from the Constitution.
- ∴ The President did not have the legal authority to seize the steel mills.

Now we proceed as before, identifying first the atomic propositions and then formalizing the argument:

³*Youngstown Sheet & Tube Co. v. Sawyer*, 343 U.S. 579 (1952).

Dictionary

A = The President had the legal authority to seize the steel mills.

C = The authority came from Congress.

O = The authority came from the Constitution.

$$(1) A \rightarrow (C \vee O)$$

$$(2) \neg C$$

$$(3) \neg O$$

$$\therefore \neg A$$

Finally, to check for validity we use a combined truth-table:

Table 3.3: Truth table for the steel-seizure argument

A	C	O	Premise 1 $A \rightarrow (C \vee O)$	Premise 2 $\neg C$	Premise 3 $\neg O$	Conclusion $\neg A$
T	T	T	T	F	F	F
T	T	F	T	F	T	F
T	F	T	T	T	F	F
T	F	F	F	T	T	F
F	T	T	T	F	F	T
F	T	F	T	F	T	T
F	F	T	T	T	F	T
F	F	F	T	T	T	T

[Try it in the Logic Lab ↗](#)

The only row where all the premises are true is the last one, and in that row the conclusion is true as well. Therefore, the argument is valid.

Exercises

Use truth-tables to show whether the following arguments are tautologically valid or invalid.

1. $A \vee B, A \rightarrow C, \neg B \vee C \therefore C$ [Practice in the Logic Lab ↗](#)

2. $P \vee (Q \wedge R) \therefore (P \vee Q) \wedge (P \vee R)$ [Practice in the Logic Lab ↗](#)

- 3. $F \rightarrow G, G \therefore F$ Practice in the Logic Lab ↗
- 4. $A, (A \wedge B) \rightarrow C \therefore B \rightarrow C$ Practice in the Logic Lab ↗
- 5. $\neg P \therefore \neg(P \vee Q)$ Practice in the Logic Lab ↗
- 6. $F, G \vee H \therefore (F \wedge G) \vee (F \wedge H)$ Practice in the Logic Lab ↗
- 7. $\neg(A \vee B) \therefore \neg A$ Practice in the Logic Lab ↗
- 8. $P \leftrightarrow Q \therefore (P \wedge Q) \vee (P \wedge \neg Q)$ Practice in the Logic Lab ↗

Translate the following natural English arguments into PL and use a truth table to show whether they are tautologically valid or invalid. Practice translating in the Logic Lab ↗

- 9. Jasmine will take a math class if she is interested in math, but she won't take a math class if she's a history major. So Jasmine isn't a history major if she's interested in math.
- 10. Curtis will get into Juilliard if he is a great dancer, and only if he practices will he be a great dancer. Therefore Curtis will get into Juilliard if he practices.
- 11. Luke must not have paid attention in his *Psychology of Intimate Relationships* class. If he paid attention, then his relationship didn't end, but it did end.
- 12. Aliyah must be hard working if she gets good grades, because she'll get good grades if she is clever or hard working.
- 13. Ryan Gosling is not famous. After all, if he won an Academy Award, then he is famous, but he has never won one.

Part II

Probability

4 Measures and stochastic truth-tables

This chapter introduces the mathematical theory of probability as a branch of measure theory. Stochastic truth-tables are used.

4.1 Introduction

Consider the following argument:

- (1) Leona has a heavy, shaggy coat, a long bushy tail, and large tufted ears.
- (2) Most Maine Coons have those characteristics, and relatively few cats which are not Maine Coons do.

∴ Leona is a Maine Coon.

That argument is not valid. However, it seems pretty good at first sight. It all depends on what we mean by "most" and "relatively few". Just as **PL** offers us a mathematical model of the truth-functional aspects of English, the theory of probability offers us a mathematical model of our ordinary talk in terms of "probable", "likely", and so on. Our task in this part is to introduce the theory of probability and give you the tools necessary to analyze arguments that are explicitly probabilistic.

We said that the preceding argument seems pretty good at first sight. But there is a crucial fact missing from the premises: the fact that under 10% of all cats are Maine Coons.¹ Suppose we add that fact to our argument:

- (1) Leona has a heavy, shaggy coat, a long bushy tail, and large tufted ears.

¹The true proportion is actually way lower than that, but we'll use the 10% figure to simplify the arguments later on.

(2) Most Maine Coons have those characteristics, and relatively few cats which are not Maine Coons do.

(3) Fewer than 1% of all cats are Maine Coons.

$\therefore$ Leona is a Maine Coon.

Should we now think that the argument is good? By the end of this part, you will be able to answer questions of this sort. In this chapter, we introduce the mathematical theory of probability as a branch of measure theory.

4.2 Measures

Probability can be defined as **a normalized measure over possibility space**. That is a mouthful, so we start by talking about measures.

A *measure*, in math, is a kind of function. We are already familiar with truth-functions: rules that take you from one (or a pair of) truth-value(s) to another truth-value. As we saw, truth-functions can be described in tables. Functions more generally are rules from something (the *domain* of the function) to something else (the *codomain* of the function). In the case of a truth-function, the domain and the codomain are both truth-values. Measures are a different kind of function. The domain of a measure is a *space* (which, as we will see, can be a literal geographical space or a more abstract notion of space as simply a set), and the codomain is the set of *non-negative real numbers*. Moreover, for a function to be a measure it has to be *additive*: if a region of the space can be completely decomposed into non-overlapping regions, then the measure of the total region must be the sum of the measures of the non-overlapping regions that compose it.

Let us start with an example: the measure whose domain is any region of New York City and whose codomain is the population of that region. We can illustrate part of that measure with a table:

Borough	Population
Brooklyn	2,500,000
Queens	2,000,000
Manhattan	1,500,000
The Bronx	1,000,000
Staten Island	500,000

We are representing just part of the function here. For instance, we do not have explicitly represented the value for the portion of Long Island that overlaps NYC. Given that that portion is just Brooklyn plus Queens, its population just is the population of Brooklyn plus the population of Queens. So, by assuming that the table above is a measure, we can extract from the table the information about the value of the function applied to that part of Long Island: 4,500,000.

Now contrast that table with the following one, which give you the proportion of the population who are Mets fans:

Borough	Mets Proportion
Brooklyn	0.6
Queens	0.75
Manhattan	0.5
The Bronx	0.25
Staten Island	0.5

That function, which takes regions of NYC and returns the proportion of the population in that region which are Mets fans, is not a measure. It satisfies the non-negativity part, but it doesn't satisfy the additivity part. For consider what result the function should yield for Long Island: *not* the sum of the values for Brooklyn and Queens (which would yield the absurd 1.35), but rather the population-weighted average of the proportion of Mets fans in Brooklyn and Queens, which is $\frac{2}{3}$.

So, to recap: measures are additive functions from regions of space to non-negative numbers. "Counting" functions, like the population function above, are measures. Proportion functions, like the proportion of Mets function, are not measures because they are not additive. Other measures include the length of portions of a line, the volume of sub-regions of a sphere, etc. (this is where the name for measures comes from). There are also measures where the "space" that is the domain of the function is a space only in a metaphorical sense. Consider, for instance, the function which takes each family member (and each set of members) as input and delivers as output the amount of fruit eaten by them per year:

Name	Fruit Consumption per Year (lbs)
Alice	105
Bruce	80
Chuck	150
Daria	75

The "space" in this case is the family in question, and the function is a measure

because if we want to know, for instance, how much fruit the parents (Alice and Bruce) consume in a year, the answer is just the sum of what they consume separately. Because the function is also non-negative, this is clearly a measure—the measure of how much fruit any group of members of this family consumes.

4.2.1 Normalized measures

We said we were going to think of probability as a *normalized measure over possibility space*. We already know what a measure is. But what does it mean to say that a measure is "normalized"? It simply means that we take the number associated with the whole space and then divide the result of the function by that total. For instance, the normalization of the population of NYC gives us the following table:

Borough	Share of population
Brooklyn	$\frac{1}{3}$
Queens	$\frac{4}{15}$
Manhattan	$\frac{1}{5}$
The Bronx	$\frac{2}{15}$
Staten Island	$\frac{1}{15}$

Notice that the value for the total space for any normalized measure will always be 1. In the case of the table above, the total space just is the total population of NYC. In the case of fruit consumption, the total "space" is the amount of fruit consumed by the whole family per year. Normalized measures give us a simple way of representing proportions.

4.2.2 Possibility space

A possibility space is the set of things that might happen. Suppose we are wondering when it would be a good day to go to the beach. Monday may have good beach weather—or not! On the other hand, one family member may have a doctor appointment on Monday. As you know, this gives us a space of four possibilities: good beach weather but doctor appointment, good beach weather without doctor appointment, bad beach weather with doctor appointment, bad beach weather without doctor appointment. Those four possibilities represent our possibility space. In general, n atomic propositions will give us a possibility space with 2^n elements. In other words, truth-tables represent possibility space.

4.2.3 Normalized measures over possibility space and stochastic truth-tables

We are now in a position to fully understand what it means to say that probability is a *normalized measure over possibility space*. For consider now what happens when we add an additional column to an ordinary truth-table for three variables, and in that column we put numbers, with just the following condition: they have to be non-negative real numbers that add up to 1:

A	B	C	$P()$
T	T	T	$1/1250$
T	T	F	$1/125$
T	F	T	$2/25$
T	F	F	$4/5$
F	T	T	$1/5000$
F	T	F	$1/1000$
F	F	T	$1/100$
F	F	F	$1/10$

Recall that we can think of each row of a truth-table as a possible world. What the table with the $P()$ column added does is assign a non-negative real number to each possible world. If we now assume that the $P()$ column represents an additive function, then we can figure out the value of that function for any proposition whatsoever that can be represented using the atomic propositions A , B , and C . The value of the function applied to a proposition will be the sum of the values of the possible worlds where the proposition is true. Consider, for instance, the proposition $A \rightarrow (B \vee C)$. You can verify, by drawing the truth-table for this proposition, that it is true in every row except row 4. Therefore, $P(A \rightarrow (B \vee C)) = \frac{1}{5}$. The $P()$ function is a probability function, and because probability functions are additive, we just need the values of the function for each possible world to find out the value of the function applied to any proposition.

We call a truth-table with a column for a probability function added a *stochastic truth-table*. Notice some interesting consequences of stochastic truth-tables. First, the probability of any tautology will be 1. That is because any tautology is true in every possible world, and we know that the sum of the probabilities for each possible world adds up to 1. On the other hand, given that contradictions are true in no possible world, the probability of any contradiction will be 0. Next, consider

a disjunction ($\varphi \vee \psi$) where φ and ψ are *mutually exclusive*:

Definition 4.1. Mutually exclusive: Two propositions φ and ψ are *mutually exclusive* if and only if there is no possible world in which they are both true—that is to say, if and only if there is no row of their truth-table in which both are true.²

So $(\varphi \vee \psi)$ will be true in those worlds where φ is true and in those worlds where ψ is true. If φ and ψ were not mutually exclusive, simply adding up those worlds could lead to double-counting. But because they are mutually exclusive, $P(\varphi \vee \psi) = P(\varphi) + P(\psi)$.

4.3 Conditional probability

So far we have been talking about the probability of a single proposition, however complex. But we are also interested in the *conditional* probability of one proposition *given*, or *conditional on*, another proposition. Suppose that we have a bag with 60 marbles. Each marble is either large or small, and each marble is either red or blue. We are told exactly how many marbles there are of each combination: 30 are large and red, 10 are large and blue, 15 are small and red, and 5 are small and blue. Suppose that we will draw a marble from the bag, and that each marble has the same probability of being drawn. We can ask questions about *unconditional* probability. For instance, the probability of drawing a large marble is $\frac{2}{3}$, because 40 of the 60 marbles are large; and the probability of drawing a red marble is $\frac{3}{4}$, because 45 of the 60 marbles are red. But suppose we want to know the probability that we draw a large marble, conditional on us drawing a red marble. How would we go about thinking about that? We are not asking what is the probability of any single proposition, not even of the proposition that if we draw a red marble then we draw a large one (more on that later). Rather, we are asking how the probabilities of two different propositions relate to each other. The intuitive idea behind conditional probability (applied to this case) is that we are zooming in on that region of the space of possibilities where we draw a red marble, and ask what proportion of that space is occupied by large marbles.

There is a definition of conditional probability that answers to this intuitive idea. We write *the probability of φ given ψ* like this: $P(\varphi | \psi)$, and we define it as a ratio of unconditional probabilities:

²This means that φ and ψ cannot both be distinct atomic sentence letters. Why?

Definition 4.2. Conditional probability: For any two propositions φ and ψ , provided that $P(\psi) \neq 0$, the probability of φ *given* ψ is

$$P(\varphi \mid \psi) = \frac{P(\varphi \wedge \psi)}{P(\psi)}.$$

That is to say, the probability of φ given ψ is the ratio of the probability of the conjunction of φ and ψ to the unconditional probability of ψ . A crucial proviso is that conditional probabilities are undefined if the proposition we are conditioning on has a probability of 0—because division by 0 is itself undefined. The idea of conditional probability is that we are zooming in on that region of possibility space where ψ is true, and we are asking in what proportion of that space φ is true as well.

Going back to our marbles example, we were wondering what is the probability of drawing a large marble, conditional on drawing a red one. Given our definition of conditional probability, we can figure out the answer just by applying the formula. Let L be *we draw a large marble*, and R be *we draw a red marble*. We are looking, then, for the solution to the following equation: $P(L \mid R) = \frac{P(L \wedge R)}{P(R)}$. We are given all the information we need to solve that equation in the description of the case: $P(R) = \frac{3}{4}$ and $P(L \wedge R) = \frac{1}{2}$, so $P(L \mid R) = \frac{1/2}{3/4} = \frac{2}{3}$.

We can also figure out conditional probabilities by using stochastic truth-tables. The description of the case gives us all the unconditional probabilities we need:

L	R	P()
T	T	1/2
T	F	1/6
F	T	1/4
F	F	1/12

How can we represent conditional probabilities with stochastic truth-tables? One thing to consider is that conditional probabilities are probabilities, and so when we add a column for conditional probabilities it will have to be filled with non-negative numbers that add up to 1. Now notice, first, that, for any proposition φ , $P(\varphi \mid \varphi) = 1$.³ So, if we are considering adding a column for the probabilities

³Why?

conditional on R , for instance, the rows where R is true will have to add up to 1, which means that the rows where R is false will have to be given 0:⁴

L	R	P()	P($\cdot$ R)
T	T	$1/2$	
T	F	$1/6$	0
F	T	$1/4$	
F	F	$1/12$	0

Now, how do we fill-in the two remaining rows? By preserving the ratios that they had before. Notice that the probability of row 1 was exactly twice that of row 3. In order to make the rows sum up to 1 again (a process called re-normalization), we therefore have to assign $\frac{2}{3}$ to row 1 and $\frac{1}{3}$ to row 3:

L	R	P()	P($\cdot$ R)
T	T	$1/2$	$2/3$
T	F	$1/6$	0
F	T	$1/4$	$1/3$
F	F	$1/12$	0

We can now read conditional probabilities off the table in just the same way as we read off unconditional ones: the probability of a proposition conditional on R is the sum of the values of the $P(\cdot | R)$ column in the rows where that proposition is true. For instance, L is true in rows 1 and 2, so $P(L | R) = \frac{2}{3} + 0 = \frac{2}{3}$, which is exactly the answer we obtained from the definition of conditional probability. Similarly, $\neg L$ is true in rows 3 and 4, so $Prb(\neg L | R) = \frac{1}{3} + 0 = \frac{1}{3}$, and the two conditional probabilities add up to 1, as they should.

Remember that we said before that we should not confuse conditional probabilities with unconditional ones, not even when the proposition whose unconditional probability we are talking about is a conditional proposition. Notice, for instance, that whereas $P(L | R) = \frac{2}{3}$, $P(R \rightarrow L) = \frac{3}{4}$.

⁴Why?

Exercises

Consider the following stochastic truth-table:

A	B	C	$P()$
T	T	T	$1/10$
T	T	F	$1/5$
T	F	T	0
T	F	F	$3/10$
F	T	T	$1/10$
F	T	F	$1/5$
F	F	T	0
F	F	F	$1/10$

Calculate the probabilities for the following formulas:

1. $(A \leftrightarrow B)$
2. $(C \rightarrow B)$
3. $(A \wedge C)$
4. $(\neg A \wedge C)$
5. Add a column for $P(\cdot \mid (B \leftrightarrow C))$ to the above table.
6. Is there a function $P()$ that meets all the following constraints?
 - $P(A \rightarrow (B \leftrightarrow C)) = 1$
 - $P(B) = P(\neg B)$
 - $P(C) = 2 \times P(C \wedge A)$
 - $P(B \wedge C \wedge \neg A) = \frac{1}{5}$

If not, prove that it's impossible. If so, provide a probability table and demonstrate that the resulting distribution satisfies each of the four constraints. (Hint: Start by building a probability table; then figure out what each of the constraints says about the probability values in the table; then figure out if it's possible to meet all of the constraints at once.)

5 Axioms, definitions, and theorems

This chapter presents probability axiomatically, defines probabilistic independence, and proves the basic theorems of the probability calculus, which we then use to solve probabilistic problems.

5.1 Axioms

In addition to thinking of probability in terms of stochastic truth-tables, we can also present it axiomatically. The usual axioms for probability are the following:

Non-negativity: For every proposition φ , $P(\varphi) \geq 0$

Normality: For every tautology τ , $P(\tau) = 1$

Additivity: For any two mutually exclusive propositions φ and ψ , $P(\varphi \vee \psi) = P(\varphi) + P(\psi)$

With this axiomatic presentation of probability we can now prove theorems about it. Our first theorem is that the probability of a contradiction is always 0. We know from our work with stochastic truth-tables that this is true, for a contradiction is true in no row of a truth-table, and so the sum of the rows where it is true is 0! Compare that insight stochastic truth-tables give us to the following proof of the theorem:

Theorem 5.1. If φ is a contradiction, then $P(\varphi) = 0$.

Proof. Suppose φ is a contradiction. Let ψ be an arbitrary tautology (for instance, $(A \vee \neg A)$). Given that φ is a contradiction and ψ a tautology, φ and ψ are

mutually exclusive. So, by **Additivity**, $P(\varphi \vee \psi) = P(\varphi) + P(\psi)$. Since ψ is a tautology, so is $(\varphi \vee \psi)$. Therefore, by **Normality**, $P(\psi) = P(\varphi \vee \psi) = 1$. Since $P(\varphi \vee \psi) = P(\varphi) + P(\psi)$, by substitution $1 = P(\varphi) + 1$. Subtracting 1 from both sides of that equation, $P(\varphi) = 0$. $\square$

We also reintroduce our definition of conditional probability:

Definition (4.2). Conditional probability: For any two propositions φ and ψ , provided that $P(\psi) \neq 0$,

$$P(\varphi \mid \psi) = \frac{P(\varphi \wedge \psi)}{P(\psi)}.$$

5.2 Independence

Suppose that we are betting on the outcome of the flips of a fair coin. In fact, the whole flipping procedure is fair: any given flip has a 0.5 probability of coming up heads and a 0.5 probability of coming up tails (we ignore other possibilities, such as that the coin lands on its side). Suppose that you have just seen 5 tails come up in a row. How do you think the next flip is going to come up?

If you think that the next flip is more likely to come up heads, the good news is that you are not alone—many people would think that. It is an intuitive answer: if 50% of the flips should come up heads, and so far there have only been tails, it seems as if heads are "due". The bad news is that this line of thinking is wrong. If you trust me that the flipping set-up is fair, then you know that the next flip is just as likely to come up heads as it is to come up tails. To think otherwise is to commit the "gambler's fallacy".

What exactly is the fallacy? We said that *any given flip* has a 0.5 probability of coming up heads and a 0.5 probability of coming up tails. That means the probability of any flip is not affected by the flips that came before (if any). When the probability of one proposition is not affected by that of another, we say that these propositions are (probabilistically) *independent*. We can give a precise definition of independence by appealing to conditional probabilities:

Definition 5.2. Probabilistic independence: φ is probabilistically independent from ψ iff $P(\psi) \neq 0$ and $P(\varphi \mid \psi) = P(\varphi)$.

The proviso that $P(\psi) \neq 0$ is there because, as we saw in Definition 4.2, $P(\varphi \mid \psi)$ is undefined when $P(\psi) = 0$.

Flippings of one fair coin are independent of each other, as are flippings of any number of coins. Another example of independence is *sampling with replacement*. Suppose we have an urn with 50 black balls and 50 white balls. *Sampling with replacement* consists of taking out one ball from the urn at random and then replacing it in the urn. Of course, the probability of the ball's being black is 50%. If after replacing it you then draw another ball at random, the probability of that second ball's being black is exactly the same: 50%. Indeed, even if you sample and replace 100 balls in a row and they are all black, the 101st draw is still 50% likely to be black. Of course, it may well happen that if you sample 100 black balls in a row you start to suspect that something fishy is going on: maybe a trick is being played on you, maybe you misheard the description of what is in the urn, or for some other reason the draws are not really independent of each other or random. But the point is that as long as you believe that the draws are independent of each other, past draws have no influence on future ones.

Coin flips are not just independent of each other, they are also *unbiased*: the probability of heads is the same as the probability of tails. That is what we mean when we say that a process is *fair*: the outcomes of the process have all the same probability and they are independent of each other:

Definition 5.3. Unbiased: a process is *unbiased* if and only if all the possible outcomes of the process have the same probability.

Definition 5.4. Fair: a process is *fair* iff it is unbiased and its outcomes are independent of each other.

5.3 Unconditional probability theorems

Let us now use the axioms and definitions to prove some more theorems. In each case, you can compare the discursive proof given here with your understanding of the theorem given by stochastic truth-tables.

Theorem 5.5. For any proposition φ , $P(\varphi) + P(\neg\varphi) = 1$.

Proof. Because φ and $\neg\varphi$ are mutually exclusive, by **Additivity**, $P(\varphi \vee \neg\varphi) = P(\varphi) + P(\neg\varphi)$. Since $(\varphi \vee \neg\varphi)$ is a tautology, $P(\varphi \vee \neg\varphi) = 1$. Since $P(\varphi \vee \neg\varphi) = P(\varphi) + P(\neg\varphi)$ and $P(\varphi \vee \neg\varphi) = 1$, by substitution, $P(\varphi) + P(\neg\varphi) = 1$. $\square$

Notice that you can use Theorem 5.5 to figure out the probability of any proposition φ if you know the probability of its negation $\neg\varphi$, and vice versa.

Theorem 5.6. If two propositions φ and ψ are tautologically equivalent, then $P(\varphi) = P(\psi)$.

Proof. Let φ and ψ be tautologically equivalent. Since they are, $(\varphi \vee \neg\psi)$ is a tautology and $(\varphi \wedge \neg\psi)$ is a contradiction. So, φ and $\neg\psi$ are mutually exclusive. Therefore, by **Additivity**, $P(\varphi \vee \neg\psi) = P(\varphi) + P(\neg\psi)$. By **Normality**, since $(\varphi \vee \neg\psi)$ is a tautology, $P(\varphi \vee \neg\psi) = 1$. Since $P(\varphi \vee \neg\psi) = 1$ and $P(\varphi \vee \neg\psi) = P(\varphi) + P(\neg\psi)$, by substitution $P(\varphi) + P(\neg\psi) = 1$. By Theorem 5.5, $P(\psi) + P(\neg\psi) = 1$, which rearranges to $P(\neg\psi) = 1 - P(\psi)$. Since $P(\neg\psi) = 1 - P(\psi)$ and $P(\varphi) + P(\neg\psi) = 1$, substitution yields $P(\varphi) + 1 - P(\psi) = 1$, which rearranges to $P(\varphi) = P(\psi)$. $\square$

Two more basic results will be useful. First, no probability can exceed 1:

Theorem 5.7. For any proposition φ , $0 \leq P(\varphi) \leq 1$.

Proof. By **Non-negativity**, $P(\varphi) \geq 0$. By Theorem 5.5, $P(\varphi) + P(\neg\varphi) = 1$, so $P(\varphi) = 1 - P(\neg\varphi)$. By **Non-negativity**, $P(\neg\varphi) \geq 0$, and so $P(\varphi) = 1 - P(\neg\varphi) \leq 1$. $\square$

Second, a proposition cannot be more probable than a proposition that is true in every row where it is true:

Theorem 5.8 (Monotonicity). If ψ is true in every row of the truth-table in which φ is true, then $P(\varphi) \leq P(\psi)$.

Proof. Suppose that ψ is true in every row in which φ is true. Then φ is tautologically equivalent to $(\varphi \wedge \psi)$, and so, by Theorem 5.6, $P(\varphi) = P(\varphi \wedge \psi)$. As a matter of logic, ψ is equivalent to $(\varphi \wedge \psi) \vee (\neg\varphi \wedge \psi)$, and $(\varphi \wedge \psi)$ and $(\neg\varphi \wedge \psi)$ are mutually exclusive. So, by Theorem 5.6 and **Additivity**, $P(\psi) = P(\varphi \wedge \psi) + P(\neg\varphi \wedge \psi)$. By **Non-negativity**, $P(\neg\varphi \wedge \psi) \geq 0$. Therefore $P(\psi) = P(\varphi) + P(\neg\varphi \wedge \psi) \geq P(\varphi)$. $\square$

In particular, $P(\varphi \wedge \psi) \leq P(\varphi)$ and $P(\varphi) \leq P(\varphi \vee \psi)$, for any propositions φ and ψ .

The **Additivity** axiom tells us how to find out the probability of a disjunction when the two disjuncts are mutually exclusive. But what if we have larger disjunctions, like $(A \vee B \vee C)$? And what if the disjuncts are not mutually exclusive? Let's tackle these questions one at a time.

Suppose we have a set of pairwise mutually exclusive propositions $\{\varphi_1, \varphi_2, \dots, \varphi_n\}$. Then:

Theorem 5.9 (extended additivity). If $\{\varphi_1, \varphi_2, \dots, \varphi_n\}$ is a set of pairwise mutually exclusive propositions, then $P(\varphi_1 \vee \varphi_2 \vee \dots \vee \varphi_n) = P(\varphi_1) + P(\varphi_2) + \dots + P(\varphi_n)$.

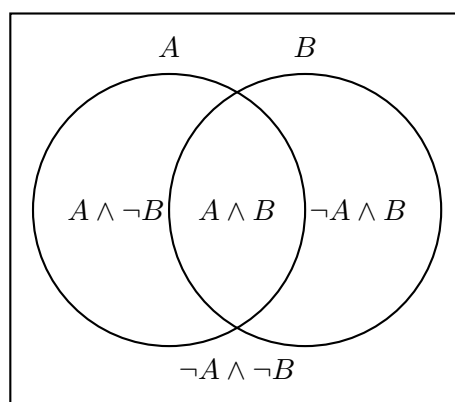
Proof. Assume that $\{\varphi_1, \varphi_2, \dots, \varphi_n\}$ is a set of pairwise mutually exclusive propositions. Then, by finite additivity, $P(\varphi_1 \vee \varphi_2) = P(\varphi_1) + P(\varphi_2)$. As a matter of logic, since φ_1 and φ_2 are each mutually exclusive with φ_3 , then so too is $(\varphi_1 \vee \varphi_2)$ mutually exclusive with φ_3 . Therefore, by another application of additivity, $P((\varphi_1 \vee \varphi_2) \vee \varphi_3) = P(\varphi_1 \vee \varphi_2) + P(\varphi_3)$. Since $P(\varphi_1 \vee \varphi_2) = P(\varphi_1) + P(\varphi_2)$ and $P((\varphi_1 \vee \varphi_2) \vee \varphi_3) = P(\varphi_1 \vee \varphi_2) + P(\varphi_3)$, substitution yields $P((\varphi_1 \vee \varphi_2) \vee \varphi_3) = P(\varphi_1) + P(\varphi_2) + P(\varphi_3)$. We can iterate this process for each φ_i to reach $P(\varphi_1 \vee \varphi_2 \vee \dots \vee \varphi_n) = P(\varphi_1) + P(\varphi_2) + \dots + P(\varphi_n)$. $\square$

What Theorem 5.9 tells us is that additivity works not just for disjunctions of two mutually exclusive disjuncts, but for disjunctions of any finite number of mutually exclusive disjuncts.

But let us now go back to our other question about **Additivity**: what if the disjuncts are not mutually exclusive? How do we find out, for instance, the probability of a disjunction of atomic propositions (or molecular propositions which are not mutually exclusive)? It would help to visualize the space of possibilities. One way of visualizing it which we are familiar with is a truth-table, where each row represents a possible world:

A	B
T	T
T	F
F	T
F	F

We can also graphically represent the same space of possibilities using a *Venn Diagram*:



We can see both in the truth-table and the Venn diagram that if we calculate $P(A \vee B)$ by adding up $P(A)$ and $P(B)$ we would be double-counting $P(A \cap B)$. Rather, what we want to do is to add the first row of the truth-table (equivalently, the intersection of A and B) just once, so that we end up with the probability of the region of the space of possibilities where *either* A or B are true:

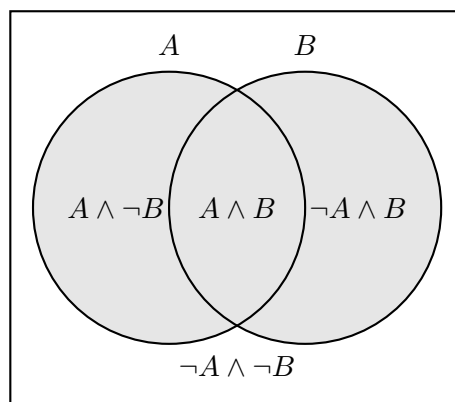


Figure 5.1: The shaded region is $A \vee B$.

So, if we know $P(\varphi)$ and $P(\psi)$, but we don't know whether φ and ψ are mutually exclusive or not, how do we determine $P(\varphi \vee \psi)$? Well, if just adding $P(\varphi)$ and $P(\psi)$ counts $P(\varphi \cap \psi)$ twice, we can just subtract it once, and so we have the following generalization of **Additivity**: $P(\varphi \vee \psi) = P(\varphi) + P(\psi) - P(\varphi \cap \psi)$, which for obvious reasons is sometimes called the *inclusion-exclusion* theorem:

Theorem 5.10 (Inclusion-exclusion theorem). For any two propositions φ and ψ , $P(\varphi \vee \psi) = P(\varphi) + P(\psi) - P(\varphi \cap \psi)$.

Proof. Left as an exercise for the reader. □

5.4 Conditional probability theorems

The inclusion-exclusion theorem determines $P(\varphi \vee \psi)$ on the basis of three other quantities: $P(\varphi)$, $P(\psi)$, and $P(\varphi \wedge \psi)$. But, assuming we are given $P(\varphi)$ and $P(\psi)$, how do we determine $P(\varphi \wedge \psi)$? The next two theorems answer that question:

Theorem 5.11. For any two events φ and ψ :

- (a) if $P(\psi) \neq 0$, then $P(\varphi \wedge \psi) = P(\varphi \mid \psi) \times P(\psi)$;
- (b) if $P(\psi) = 0$, then $P(\varphi \wedge \psi) = 0$.

Proof. There are two possibilities— $P(\psi) \neq 0$ or $P(\psi) = 0$ —that we will consider in turn.

- (a) If $P(\psi) \neq 0$, then $P(\varphi \wedge \psi) = P(\varphi \mid \psi) \times P(\psi)$.

Proof. Suppose that $P(\psi) \neq 0$. By Definition 4.2, since $P(\psi) \neq 0$, $P(\varphi \mid \psi) = \frac{P(\varphi \wedge \psi)}{P(\psi)}$, which rearranges to $P(\varphi \wedge \psi) = P(\varphi \mid \psi) \times P(\psi)$. $\square$

- (b) If $P(\psi) = 0$, then $P(\varphi \wedge \psi) = 0$.

Proof. Suppose that $P(\psi) = 0$. As a matter of logic, ψ is equivalent to $(\varphi \wedge \psi) \vee (\neg\varphi \wedge \psi)$. By Theorem 5.6, since ψ is equivalent to $(\varphi \wedge \psi) \vee (\neg\varphi \wedge \psi)$, $P(\psi) = P((\varphi \wedge \psi) \vee (\neg\varphi \wedge \psi))$. $(\varphi \wedge \psi)$ and $(\neg\varphi \wedge \psi)$ are mutually exclusive, and so by **Additivity**, since $(\varphi \wedge \psi)$ and $(\neg\varphi \wedge \psi)$ are mutually exclusive, $P((\varphi \wedge \psi) \vee (\neg\varphi \wedge \psi)) = P(\varphi \wedge \psi) + P(\neg\varphi \wedge \psi)$. Thus, since $P(\psi) = 0$ and $P(\psi) = P((\varphi \wedge \psi) \vee (\neg\varphi \wedge \psi))$, $P(\varphi \wedge \psi) + P(\neg\varphi \wedge \psi) = 0$. By **Non-negativity**, neither $P(\varphi \wedge \psi)$ nor $P(\neg\varphi \wedge \psi)$ can be negative. Thus, since $P(\varphi \wedge \psi) + P(\neg\varphi \wedge \psi) = 0$, it must be that $P(\varphi \wedge \psi) = P(\neg\varphi \wedge \psi) = 0$. $\square$

Since one of $P(\psi) \neq 0$ or $P(\psi) = 0$ must be true, (a) and (b) together establish the theorem. (When $P(\psi) = 0$ the conditional probability $P(\varphi \mid \psi)$ is undefined, which is why the first claim of the theorem is restricted to the case $P(\psi) \neq 0$.) $\square$

As a special case of Theorem 5.11, we can prove that when φ is independent of ψ , $P(\varphi \wedge \psi) = P(\varphi) \times P(\psi)$:

Theorem 5.12. If an event φ is independent of an event ψ , then $P(\varphi \wedge \psi) = P(\varphi) \times P(\psi)$.

Proof. Let φ be independent of ψ . By Definition 5.2, since φ is independent of ψ , $P(\varphi) = P(\varphi | \psi)$. By Definition 5.2, $P(\psi) \neq 0$, and so by Theorem 5.11(a), $P(\varphi \wedge \psi) = P(\varphi | \psi) \times P(\psi)$. By substituting $P(\varphi)$ in for $P(\varphi | \psi)$, we get $P(\varphi \wedge \psi) = P(\varphi) \times P(\psi)$. $\square$

Going back to our discussion of the Gambler's fallacy, Theorem 5.12 tells us that the probability of getting six tails in a row is exactly the same as the probability of getting five tails followed by heads, namely $0.5^6 = \frac{1}{64}$. To see why, let T_i be the proposition that the i th flip comes up tails. When we say that the flips of a fair coin are independent of each other, we mean that each flip is independent not just of each earlier flip taken separately, but of any combination of the earlier flips. So the sixth flip is independent of $T_1 \wedge \dots \wedge T_5$, and by Theorem 5.12, $P(T_1 \wedge \dots \wedge T_6) = P(T_1 \wedge \dots \wedge T_5) \times P(T_6)$. Applying the same reasoning to the earlier flips, $P(T_1 \wedge \dots \wedge T_6) = 0.5^6$. The same goes for five tails followed by heads. This is just what the Gambler's fallacy overlooks: the conditional probability of heads on the sixth flip, given that the previous five were tails, is the same as the conditional probability of tails on the sixth flip, namely 0.5, because every flip is independent of the previous ones.

Ignoring cases where one of the propositions has probability 0 (which can give rise to strange independence relations), probabilistic independence is symmetric:

Theorem 5.13. For any two events φ and ψ , if $P(\varphi) \neq 0$ and $P(\psi) \neq 0$, then φ is independent of ψ if and only if ψ is independent of φ .

Proof. Let $P(\varphi) \neq 0$ and $P(\psi) \neq 0$. We prove the biconditional—that φ is independent of ψ if and only if ψ is independent of φ —by proving the two conditionals that constitute it.

(a) φ is independent of ψ only if ψ is independent of φ .

Proof. Suppose that φ is independent of ψ . By Theorem 5.12, since φ is independent of ψ , $P(\varphi \wedge \psi) = P(\varphi) \times P(\psi)$. Since $P(\varphi) \neq 0$ and $P(\varphi \wedge \psi) = P(\varphi) \times P(\psi)$, $P(\psi) = \frac{P(\varphi \wedge \psi)}{P(\varphi)}$. By Definition 4.2, since $P(\varphi) \neq 0$, $P(\psi | \varphi) = \frac{P(\psi \wedge \varphi)}{P(\varphi)}$. By Theorem 5.6 and the logical equivalence of $\psi \wedge \varphi$ and $\varphi \wedge \psi$, $P(\psi \wedge \varphi) = P(\varphi \wedge \psi)$. Since $P(\psi \wedge \varphi) = P(\varphi \wedge \psi)$ and $P(\psi | \varphi) = \frac{P(\psi \wedge \varphi)}{P(\varphi)}$, $P(\psi | \varphi) = \frac{P(\varphi \wedge \psi)}{P(\varphi)}$. Since $P(\psi) = \frac{P(\varphi \wedge \psi)}{P(\varphi)}$ and $P(\psi | \varphi) = \frac{P(\varphi \wedge \psi)}{P(\varphi)}$, $P(\psi) = P(\psi | \varphi)$. By Definition 5.2, since $P(\varphi) \neq 0$ and $P(\psi) = P(\psi | \varphi)$, ψ is independent of φ . $\square$

(b) φ is independent of ψ if ψ is independent of φ .

Proof. Exactly as the proof of (a) above, but with every instance of φ replaced with ψ , and every instance of ψ replaced with φ . $\square$

By (a) and (b), we have it that φ is independent of ψ if and only if ψ is independent of φ . $\square$

We say that φ and ψ are *independent of each other* (or just *independent*) when each of them is independent of the other. By Theorem 5.13, for propositions with non-zero probability this is the same as either one of them being independent of the other.

If a proposition φ has probability 0, then it is independent of any proposition ψ that has non-zero probability—see exercises below. But if $P(\varphi) = 0$, then the probability of any proposition conditional on φ is undefined—see Definition 4.2—and so, by the definition of independence, no proposition at all is independent of φ .

5.5 Solving problems

Armed with the axioms and theorems of the probability calculus, we can now solve a number of probabilistic problems. Let's try a sample one:

Assume: $P(H) = \frac{1}{2}$, $P(G) = \frac{1}{5}$, $P(\neg F) = \frac{3}{5}$, $P(H \vee G) = \frac{3}{5}$, $P(H|F) = \frac{1}{2}$,
and F and G are independent of each other.

What is $P(H \wedge G)$?

The basic strategy for solving any problem like this is to:

1. Check what your goal is and note what's involved in it.
2. Check what you already have and write down anything that involves the same elements as your goal.
3. Write down any axioms, theorems, and definitions that pertain to the kinds of things you are looking for or already have.
4. Check if you have everything written down that you need in order to solve for your goal.
If you do, just solve for your goal and you're done!
If you don't...

5. Repeat from step 1 with what you're missing as your new goal.

Let's apply this procedure to our sample problem. Our goal is to find out $P(H \wedge G)$, which involves H, G , and conjunction. The second step is to check what we already have, which in this case is a lot. What is particularly relevant to our goal are the following:

- $P(H) = \frac{1}{2}$
- $P(G) = \frac{1}{5}$
- $P(H \vee G) = \frac{3}{5}$
- $P(H \mid F) = \frac{1}{2}$
- F and G are independent.

The third step is to examine the axioms, theorems and definitions to find if anything there is relevant to our goal.¹ Definition 4.2 concerns conjunction:

Definition (4.2). Conditional Probability: For any two propositions φ and ψ , provided that $P(\psi) \neq 0$,

$$P(\varphi \mid \psi) = \frac{P(\varphi \wedge \psi)}{P(\psi)}$$

φ and ψ are metavariables ranging over propositions of **PL**. What we are interested in is the conjunction of H and G , so we should apply Definition 4.2 to those propositions:

Provided that $P(G) \neq 0$,

$$P(H \mid G) = \frac{P(H \wedge G)}{P(G)}$$

We are also given that $P(G) = \frac{1}{5}$, so that the proviso of Definition 4.2 is satisfied. In order to solve for our goal— $P(H \wedge G)$ —we need to know the values of $P(H \mid G)$ and $P(G)$. We know that $P(G) = \frac{1}{5}$, but we don't know $P(H \mid G)$. Maybe we'll have to find $P(H \mid G)$ as an intermediate step in solving this problem, so let's not completely abandon this line of reasoning, but we can set it aside for now to see if there is some easier way to do it.

¹Remember that all axioms, theorems and definitions are listed at the end of this book.

Going back to our list of axioms and theorems that are relevant to our goal we find Theorem 5.10, according to which $P(\varphi \vee \psi) = P(\varphi) + P(\psi) - P(\varphi \wedge \psi)$. Instantiating this theorem to G and H gives us:

$$P(H \vee G) = P(H) + P(G) - P(H \wedge G)$$

Does this help? It does, because we know that $P(H \vee G) = \frac{3}{5}$, $P(H) = \frac{1}{2}$, and $P(G) = \frac{1}{5}$. Substituting in those values in the above formula gives us:

$$\frac{3}{5} = \frac{1}{2} + \frac{1}{5} - P(H \wedge G)$$

Adding $P(H \wedge G)$ and subtracting $\frac{3}{5}$ from both sides gives us our solution:

$$P(H \wedge G) = \frac{1}{2} + \frac{1}{5} - \frac{3}{5} = \frac{1}{2} - \frac{2}{5} = \frac{5}{10} - \frac{4}{10} = \frac{1}{10}$$

Exercises

1. Suppose you are going to draw a card at random from a standard deck. For each pair of propositions, say whether they are independent or not.

- a) A = the card is red.
 B = the card is an ace.
- b) A = the card is red.
 B = the card is a diamond.
- c) A = the card is an ace.
 B = the card is a spade.
- d) A = the card is a Queen.
 B = the card is a face card.

2. For each of the following examples, say whether it's an instance of the Gambler's fallacy

- a) You're playing cards with your friends using a standard, randomly shuffled deck of 52 cards. You're about half-way through the deck and no aces have been drawn yet. You conclude that an ace is due soon, and thus the probability the next card is an ace has gone up.

- b) You're holding a six-sided die, which you know to be fair. You're going to roll it 60 times. You figure about 10 of those rolls should be threes. But after 59 rolls, you've rolled a three only five times. You figure that the probability of a three on the last roll has gone up: it's higher than just $\frac{1}{6}$.
 - c) You know the lottery numbers in New Jersey are selected using a fair process. So it's really unlikely that someone will win two weeks in a row. Your friend won last week, so you conclude their chances of winning this week too are even lower than usual.
 - d) You're visiting a new country where corruption is common, so you aren't sure whether the lottery there is fair. You see on the news that the King's cousin won the lottery two weeks in a row. You conclude that their chances of winning next week are higher than normal, because the lottery is rigged in their favor.
- 3. Prove Theorem 5.10.**
- 4. Prove that if φ has probability 0, then it is independent of any proposition ψ with $P(\psi) \neq 0$. (Hint: by Theorem 5.8, $P(\varphi \wedge \psi) \leq P(\varphi)$.)**
- 5. Say whether the following event pairs are independent with each other. (All when rolling a pair of dice...)**
- a) "The sum of the faces is 5." and "The first die comes up greater than 3, and the second less than 4."
 - b) "The sum of the faces is 5." and "The second die comes up less than 4."
 - c) "The sum of the faces is 7." and "The second die comes up odd."
 - d) "The sum of the faces is 4." and "The sum of the faces is 10."
 - e) "The sum of the faces is 7 and the first die comes up odd." and "The first die comes up greater than 2."
- 6. Use the suppositional information to find the unknown probability.**
- a) Suppose that J and K are independent and $P(K \mid J) = 3/4$. What is $P(K)$?
 - b) Suppose that F and G are independent and $P(F) = 5/9$. What is $P(F \mid G)$?

- c) What is $P(H \vee \neg H)$?
 - d) Suppose that F and G are mutually exclusive, $P(F) = 1/4$, and $P(G) = 2/3$. What is $P(F \vee G)$?
 - e) Suppose that J and K are mutually exclusive, $P(J \vee K) = 3/8$, and $P(J) = 1/4$. What is $P(K)$?
 - f) Suppose that $P(X) = 2/3$ and $P(X \wedge Y) = 3/7$. What is $P(Y | X)$?
 - g) Suppose that $P(F \wedge G) = 2/5$ and $P(G | F) = 3/4$. What is $P(F)$?
 - h) What is $P(H \wedge \neg H)$?
 - i) Suppose that F and G are mutually exclusive. What is $P(G \wedge F)$?
 - j) Suppose that $P(J) = 4/7$. What is $P(\neg J)$?
 - k) Suppose that $P(R) = 1/3$ and R and S are logically equivalent. What is $P(S)$?
 - l) Suppose that $P(G) = 1/3$ and F and G are logically equivalent. What is $P(\neg F)$?
 - m) Suppose that $P(J) = 5/8$, $P(K) = 3/4$, and $P(J \wedge K) = 3/8$. What is $P(J \vee K)$?
 - n) Suppose that $P(M) = 2/3$, $P(L \wedge M) = 1/2$, and L is independent of M . What is $P(L)$?
 - o) Suppose that $P(S | R) = 1/3$ and $P(S) = 1/3$. Is R independent of S ?
 - p) Suppose that $P(G \wedge H) = 1/3$, $P(G) = 1/2$, and $P(H) = 1/3$. Is H independent of G ?
 - q) Suppose that $P(J | K) = 1/2$, $P(J) = 1/3$, and $P(K) = 3/5$. What is $P(K | J)$?
 - r) Suppose that $P(L | M) = 5/9$, $P(M | L) = 2/3$, and $P(M) = 3/5$. What is $P(L)$?
- 7. Let the following be true:** $P(H) = 1/2$, $P(G) = 1/5$, $P(\neg F) = 3/5$, $P(H \vee G) = 3/5$, $P(H | F) = 1/2$, F is independent of G . **Find the values of the probabilities below.**
- a) $P(F)$
 - b) $P(F \wedge G)$
 - c) $P(F \wedge H)$

d) $P(G | F)$

e) $P(F | H)$

- 8. Let the following be true:** $P(J) = 3/5$, $P(K) = 2/5$, $P(L \wedge K) = 3/10$, $P(L \vee K) = 3/5$, $P(L | J) = 2/3$, K is logically equivalent to $\neg M$, and J is independent of M . Find the values of the probabilities below.

a) $P(M)$

b) $P(J | M)$

c) $P(J | L)$

d) $P(K \vee M)$

e) $P(J \wedge L)$

- 9. Let the following be true:** $P(X) = 1/2$, $P(Y) = 1/2$, $P(X \wedge Y) = 1/4$.

a) Are X and Y mutually exclusive?

b) What is $P(X \vee Y)$?

6 Bayesianism

This chapter motivates credences with the lottery paradox and introduces Bayesianism: the view that credences should satisfy the probability axioms and should be updated by conditionalizing on new evidence. We prove Bayes' theorem, use it (and eikosograms) to compute posterior credences, apply it to Hume's argument about miracles, and close with the problem of priors.

6.1 The lottery paradox: beliefs and credences

In order to win Italy's "SuperEnalotto" lottery, one has to correctly predict 6 numbers out of a pool of 90 (order doesn't matter). There are 622,614,630 ways of picking 6 numbers out of 90.¹ Consider all those possible entries in the lottery, one by one, and suppose that for each one we ask: is it going to win? If forced to answer "Yes" or "No", I would say "No". And it is also natural to think that buying an entry in the SuperEnalotto is a waste of money because one's entry is not going to win. This suggests that we believe, of any given entry, that it is not going to win. Fix some way of numbering the possible entries (for instance, order the numbers from smallest to largest in each entry, and order the entries from smallest to largest resulting number). Let "1" represent the proposition *ticket number 1 will win*, "2" the proposition *ticket number 2 will win*; in general, let n represent the proposition *ticket number n will win*. If we believe, of each ticket, that it will not win, then believe all of these propositions:

¬1
¬2
...
¬622,614,630

¹The general formula to calculate how many ways there are to pick r things out of n possibilities is: ${}_nC_r = \frac{n!}{r!(n-r)!}$.

But we also believe that one of those tickets will win. That is to say, we believe the following disjunction:

$$1 \vee 2 \vee \dots \vee 622,614,630$$

Those 622,614,631 beliefs of ours are contradictory. This is easy to see: if each ticket is a loser (as our first 622,614,630 beliefs say they are), then no ticket is a winner (contradicting our 622,614,631st belief). Apparently, the intuitive reaction to what we should believe about a lottery lands us in contradiction.

Some think that the way out of this contradiction is to realize that not all of our opinions are all-or-nothing. The lottery paradox arises because we are forced to say whether we believe that a ticket will win or not. But perhaps our opinion is not well represented by a "Yes" or "No" answer to that question, but rather by a graded answer: we are pretty sure that each ticket is going to lose, but we are not certain. In fact, we can even quantify how sure we are, of each ticket, that it is going to lose: we are $1 - \frac{1}{622,614,630}$ sure, of each ticket, that it is going to lose—and $\frac{1}{622,614,630}$ sure that it is going to win. These *degrees of beliefs* are sometimes called *credences*. Once we think of the SuperEnalotto in terms of credences instead of all-or-nothing beliefs, the paradox disappears. Notice that $\{1, 2, \dots, 622,614,630\}$ is a set of pairwise mutually exclusive propositions (one ticket's winning is incompatible with any other ticket's winning: we stipulate that the possible worlds under consideration are just those in which exactly one ticket wins). Therefore, by Theorem 5.9, the probability that any one of a number of tickets will win is the sum of the probability that each one of them will win. Therefore, the probability that one ticket will win is 1! If we align our credences with these (known to us) probabilities, then we can have, for each ticket, an extremely low credence that it will win, while at the same time having credence 1 that some ticket will win.

Needless to say, not all philosophers are convinced by the above argument that credences are psychologically real. But in what follows we proceed with the hypothesis that they are, and see where that leads us.

6.2 Bayesianism

Our credence-first treatment of the lottery paradox had us align our credences with the known probabilities, and thus our credences themselves were probabilities: they satisfied **Non-negativity**, **Normality**, and **Additivity**. Some people believe that this should always be the case: they think that, in order to be rational, our

credences should satisfy the probability axioms. Because it applies to our credences *at a time*, we will call this the *synchronic norm for credences*:

Norm 6.1. Synchronic norm for credences: Credences at a time must be probabilistically coherent (i.e., they must obey the probability axioms).

The **synchronic norm for credences** establishes relationships between the credences we have in different propositions at a time. For instance, our credences in a proposition and its negation cannot float free from each other—one must be 1 – the other.

But we are not just interested in how our credences relate to each other at a single time, we also want to know how to change our credences once we acquire new evidence. To simplify discussion, let us concentrate on just two propositions H and E . Let $E = \text{Shaun studies hard for the exam}$, and $H = \text{Shaun passes the exam}$. How should we fill-in the stochastic truth-table for these two propositions? To do so we need to assign credences not just to the individual propositions H and E , but to each of the four possible worlds determined by them—to each row of the stochastic truth-table. Thus, let us suppose that we are 50-50 on whether Shaun will pass the exam, when that proposition is considered independently of whether or not he studies hard, and let us also suppose that we are 50-50 on whether or not he will study hard. However, we are much more confident that he will pass conditional on his studying than otherwise—let us say, we think the possible world where he passes and studies hard is 9 times more likely than the possible world where he passes but doesn't study hard. In that case, we will fill-in the stochastic truth-table as follows (notice that we replace the name of the P column to Cr , because we are thinking of credences now):

H	E	Cr
T	T	9/20
T	F	1/20
F	T	1/20
F	F	9/20

Notice that, as we know, unconditional probabilities determine conditional probabilities. In particular, we can add a column for credences conditional on E : As stipulated, our credence that Shaun will pass the test, conditional on his studying hard, is 0.9, whereas our credence that he will pass conditional on his not studying is only 0.1 (the world where he passes without studying has credence $1/20$, which is one tenth of our credence of $1/2$ that he does not study). Notice that all of these

H	E	Cr	Cr($\cdot$ E)
T	T	9/20	9/10
T	F	1/20	0
F	T	1/20	1/10
F	F	9/20	0

credences—conditional and unconditional—do not commit us one way or the other on either whether Shaun will pass the exam or whether he will study—we are still 50-50 on those propositions.

But now suppose that we learn that Shaun did study hard. How should we change our credences? Should we still be 50-50 on whether he will pass? Well, should we still be 50-50 on whether he studied hard? Clearly not: we learned that he did study hard, so we should now have credence 1 that he did and credence 0 that he did not. And how should that affect our credence that he will pass? If you think it should make us more confident that he will pass, you are not alone. Indeed, many philosophers have argued that we should now be 90% confident that he will pass the exam. Those philosophers adhere to the following *diachronic norm for credences*:

Norm 6.2. Diachronic norm for credences: When E is the total evidence you learn, you should change your credences by *conditionalizing on E* . That is to say (where Cr_{old} is your credence function before learning E and Cr_{new} is your credence function after learning E):

For all φ ,

$$\text{Cr}_{new}(\varphi) = \text{Cr}_{old}(\varphi \mid E).$$

We have to be very careful in distinguishing the definition of conditional probability from the **diachronic norm for credences**. The former is just stipulated to be a ratio of your unconditional credences at a time, whereas the latter is a norm for how your credences should change over time. Of course, the **diachronic norm for credences** is intimately related to conditional probabilities, for it commands you to change your credences by appealing to your old conditional credences—but they are still two very different things. For instance, many would reject the **diachronic norm for credences**, but that doesn't commit them to denying the existence of conditional probabilities.

"Bayesianism" will be our name for the view that adopts both the synchronic and the diachronic norms for credences. In order to explain the name, we need to introduce one more theorem about conditional probabilities:

Theorem 6.3 (Bayes' theorem). For any φ and ψ , if $P(\varphi) \neq 0$ and $P(\psi) \neq 0$, then

$$P(\varphi | \psi) = \frac{P(\psi | \varphi) \times P(\varphi)}{P(\psi)}.$$

Proof. Let $P(\varphi) \neq 0$ and $P(\psi) \neq 0$. By Definition 4.2, since $P(\psi) \neq 0$, $P(\varphi | \psi) = \frac{P(\varphi \wedge \psi)}{P(\psi)}$. By Theorem 5.11(a), since $P(\varphi) \neq 0$, $P(\psi \wedge \varphi) = P(\psi | \varphi) \times P(\varphi)$. By Theorem 5.6 and the logical equivalence of $\psi \wedge \varphi$ and $\varphi \wedge \psi$, $P(\psi \wedge \varphi) = P(\varphi \wedge \psi)$. Since $P(\psi \wedge \varphi) = P(\varphi \wedge \psi)$ and $P(\psi \wedge \varphi) = P(\psi | \varphi) \times P(\varphi)$, $P(\varphi \wedge \psi) = P(\psi | \varphi) \times P(\varphi)$. Since $P(\varphi \wedge \psi) = P(\psi | \varphi) \times P(\varphi)$ and $P(\varphi | \psi) = \frac{P(\varphi \wedge \psi)}{P(\psi)}$,

$$P(\varphi | \psi) = \frac{P(\psi | \varphi) \times P(\varphi)}{P(\psi)}.$$

□

The theorem is named after Thomas Bayes, an eighteenth-century English statistician, philosopher, and Presbyterian minister who first formulated a special case of it. It relates the conditional probability of, e.g., H given E , to the converse conditional probability of E given H . The theorem should not be confused either with the definition of conditional probability or with the diachronic norm for credences. Its importance, however, is related to the diachronic norm for credences, because the information we have often makes it easier to calculate conditional probabilities by appealing to Bayes' theorem rather than the definition of conditional probability—and, of course, conditional probabilities need to be calculated in order to apply the diachronic norm for credences.

The quantities in Bayes' theorem have traditional names. $P(\varphi | \psi)$ is called the *posterior* probability, $P(\varphi)$ and $P(\psi)$ the *prior* probabilities, and $P(\psi | \varphi)$ the *likelihood* (of the hypothesis φ , relative to the evidence ψ). Let us apply the theorem to our example about Shaun and his exam:

$$\text{Cr}(H | E) = \frac{\text{Cr}(E | H) \times \text{Cr}(H)}{\text{Cr}(E)} = \frac{\frac{9}{10} \times \frac{1}{2}}{\frac{1}{2}} = \frac{9}{10}$$

$\text{Cr}(H | E)$ is called the *posterior* probability because it is the credence you will have in H **after** conditionalizing on E , and $\text{Cr}(H)$ is called the *prior* probability because it is credence you have in H **before** conditionalizing on E . Why $\text{Cr}(E | H)$ is called the *likelihood* is a mystery.²

There is another form of Bayes' theorem that will be useful as well. In order to state it, we need to prove another theorem:

²Not really—it's there to distinguish it from the probability of H given E .

Theorem 6.4 (Law of Total Probability). For any φ and ψ , if $0 < P(\psi) < 1$, then $P(\varphi) = P(\varphi \mid \psi) \times P(\psi) + P(\varphi \mid \neg\psi) \times P(\neg\psi)$.

Proof. Let $0 < P(\psi) < 1$. By Theorem 5.5, $P(\psi) + P(\neg\psi) = 1$, and so, since $P(\psi) < 1$, $P(\neg\psi) \neq 0$. By Theorem 5.11(a), since $P(\psi) \neq 0$, $P(\varphi \wedge \psi) = P(\varphi \mid \psi) \times P(\psi)$. By Theorem 5.11(a), since $P(\neg\psi) \neq 0$, $P(\varphi \wedge \neg\psi) = P(\varphi \mid \neg\psi) \times P(\neg\psi)$. As a matter of logic, φ is equivalent to $(\varphi \wedge \psi) \vee (\varphi \wedge \neg\psi)$. By Theorem 5.6, since φ is equivalent to $(\varphi \wedge \psi) \vee (\varphi \wedge \neg\psi)$, $P(\varphi) = P((\varphi \wedge \psi) \vee (\varphi \wedge \neg\psi))$. As a matter of logic, $(\varphi \wedge \psi) \wedge (\varphi \wedge \neg\psi)$ is a contradiction, and so $(\varphi \wedge \psi)$ and $(\varphi \wedge \neg\psi)$ are mutually exclusive. By **Additivity**, since $(\varphi \wedge \psi)$ and $(\varphi \wedge \neg\psi)$ are mutually exclusive, $P((\varphi \wedge \psi) \vee (\varphi \wedge \neg\psi)) = P(\varphi \wedge \psi) + P(\varphi \wedge \neg\psi)$. Since $P(\varphi) = P((\varphi \wedge \psi) \vee (\varphi \wedge \neg\psi))$ and $P((\varphi \wedge \psi) \vee (\varphi \wedge \neg\psi)) = P(\varphi \wedge \psi) + P(\varphi \wedge \neg\psi)$, $P(\varphi) = P(\varphi \wedge \psi) + P(\varphi \wedge \neg\psi)$. Since $P(\varphi \wedge \psi) = P(\varphi \mid \psi) \times P(\psi)$ and $P(\varphi \wedge \neg\psi) = P(\varphi \mid \neg\psi) \times P(\neg\psi)$ and $P(\varphi) = P(\varphi \wedge \psi) + P(\varphi \wedge \neg\psi)$, $P(\varphi) = P(\varphi \mid \psi) \times P(\psi) + P(\varphi \mid \neg\psi) \times P(\neg\psi)$. $\square$

Theorem 6.4 tells us that the probability of any proposition is an average of its probabilities conditional on any other proposition and its negation, weighted by the probabilities of that other proposition and its negation. This means that we can re-write the denominator in Bayes' theorem (notice that when we use Theorem 6.4 in the theorem below we switch φ and ψ , to match Theorem 6.3):

Theorem 6.5 (Bayes' theorem — LTP). For any φ and ψ , if $0 < P(\varphi) < 1$ and $P(\psi) \neq 0$, then

$$P(\varphi \mid \psi) = \frac{P(\psi \mid \varphi) \times P(\varphi)}{P(\psi \mid \varphi) \times P(\varphi) + P(\psi \mid \neg\varphi) \times P(\neg\varphi)}$$

Proof. Let $0 < P(\varphi) < 1$ and $P(\psi) \neq 0$. Since $P(\varphi) \neq 0$ and $P(\psi) \neq 0$, Theorem 6.3 applies, and so by Theorem 6.3,

$$P(\varphi \mid \psi) = \frac{P(\psi \mid \varphi) \times P(\varphi)}{P(\psi)}.$$

By Theorem 6.4, since $0 < P(\varphi) < 1$, $P(\psi) = P(\psi \mid \varphi) \times P(\varphi) + P(\psi \mid \neg\varphi) \times P(\neg\varphi)$. By substitution,

$$P(\varphi \mid \psi) = \frac{P(\psi \mid \varphi) \times P(\varphi)}{P(\psi \mid \varphi) \times P(\varphi) + P(\psi \mid \neg\varphi) \times P(\neg\varphi)}$$

$\square$

This version of Bayes' theorem is useful because we will sometimes not know what the probability of the evidence is directly, but we will know the likelihood of the evidence given both the hypothesis and its negation.

Let us exemplify with Leona. Recall the argument from Chapter 4:

- (1) Leona has a heavy, shaggy coat, a long bushy tail, and large tufted ears.
- (2) Most Maine Coons have those characteristics, and relatively few cats which are not Maine Coons do.

$\therefore$ Leona is a Maine Coon.

Is that argument good? It is not valid, but we know that not all good arguments are valid. Let us try to formalize it and apply our probability tools to see how good it is. We start with a dictionary. The first premise and the conclusion are easy to symbolize:

Dictionary:

L = Leona has a heavy, shaggy coat, a long bushy tail, and large tufted ears.

M = Leona is a Maine Coon.

But how should we symbolize the second premise? To begin with, we'll do well to separate its two components: most Maine coons have a heavy, shaggy coat, a long bushy tail, and large tufted ears; most non-Maine Coon cats do not. We do not want to assign to each one of these other atomic propositions, because that way we would be losing important information. Rather, we want to see how to connect that second premise to the other one and the conclusion. Intuitively, the idea is clear: any cat's being a Maine Coon must raise the probability that it has the characteristic Leona has, and any cat's not being a Maine Coon must lower it. Let us suppose that $\text{Cr}_{old}(L \mid M) = \frac{9}{10}$ and that $\text{Cr}_{old}(L \mid \neg M) = \frac{1}{5}$. Now we are told that L is true. How should we update our credences? According to the diachronic norm for credences, we should do the following:

$$\text{Cr}_{new}(M) = \text{Cr}_{old}(M \mid L)$$

We can now try to apply Bayes' theorem to this problem:

$$\text{Cr}_{old}(M \mid L) = \frac{\text{Cr}_{old}(L \mid M) \times \text{Cr}_{old}(M)}{\text{Cr}_{old}(L)} = \frac{\frac{9}{10} \times \text{Cr}_{old}(M)}{\text{Cr}_{old}(L)}$$

That wasn't too much progress, because we are left with two unknowns still. Switching to the LTP version of Bayes' theorem helps a lot:

$$\begin{aligned}\text{Cr}_{old}(M | L) &= \frac{\text{Cr}_{old}(L | M) \times \text{Cr}_{old}(M)}{\text{Cr}_{old}(L | M) \times \text{Cr}_{old}(M) + \text{Cr}_{old}(L | \neg M) \times \text{Cr}_{old}(\neg M)} \\ &= \frac{\frac{9}{10} \times \text{Cr}_{old}(M)}{\frac{9}{10} \times \text{Cr}_{old}(M) + \frac{2}{10} \times \text{Cr}_{old}(\neg M)}\end{aligned}$$

But we still have two unknowns: $\text{Cr}_{old}(M)$ and $\text{Cr}_{old}(\neg M)$. Well, actually it's just one unknown, because we know by Theorem 5.5 that one of them can be derived from the other. So, what should $\text{Cr}(M)$ be—that is to say, what should our credence that Leona is a Maine Coon be before we know anything about her? Well, remember the other fact we mentioned about Maine Coons: that they make up fewer than 10% of all cats. If we guide our priors by this, then $\text{Cr}(M) = \frac{1}{10}$ and $\text{Cr}(\neg M) = \frac{9}{10}$. Plugging in those numbers:

$$\begin{aligned}\text{Cr}_{old}(M | L) &= \frac{\frac{9}{10} \times \frac{1}{10}}{\frac{9}{10} \times \frac{1}{10} + \frac{2}{10} \times \frac{9}{10}} \\ &= \frac{\frac{9}{100}}{\frac{9}{100} + \frac{18}{100}} \\ &= \frac{\frac{9}{100}}{\frac{27}{100}} \\ &= \frac{9}{27} \\ &= \frac{1}{3}\end{aligned}$$

And so $\text{Cr}_{new}(M) = \frac{1}{3}$. Which means two things: first, the fact that Leona has Maine Coon-ish characteristics is indeed good evidence that she is a Maine Coon—we went from $\text{Cr}_{old}(M) = \frac{1}{10}$ to $\text{Cr}_{new}(M) = \frac{1}{3}$, which means that our credence that Leona is a Maine Coon more than tripled; but second, we still think that it is twice as likely that Leona is *not* a Maine Coon. How could that be? Relatedly, how could it be that having Maine Coon-ish characteristics is such good evidence for being a Maine Coon while at the same time most cats with those characteristics are not Maine Coons? The answer is in the priors. There are many more cats that are not Maine Coons than there are cats that are. Thus, suppose that we

concentrate on a random population of 100 cats. 10 of them will be Maine Coons, and of those 9 will have Maine Coon-ish characteristics. Meanwhile, 90 of them will not be Maine Coons, and 18 of those will have Maine Coon-ish features. Thus, if a randomly selected cat has Maine Coon-ish features, it is twice as likely to not be a Maine Coon as to be one—even when those features are had by 90% of Maine Coons and only by 20% of cats that are not Maine Coons.

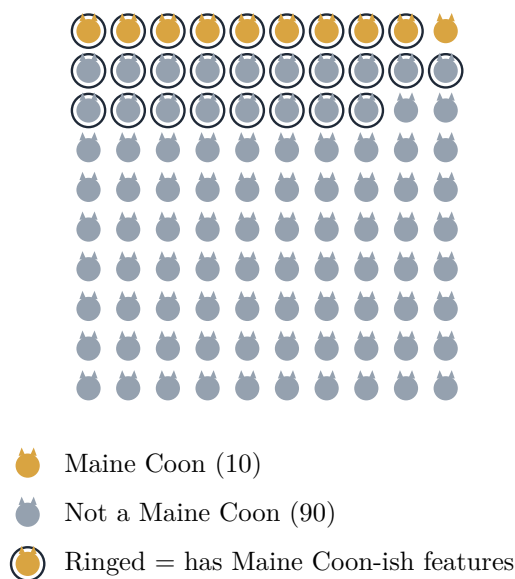


Figure 6.1: A random population of 100 cats. Of the 10 Maine Coons (gold), 9 have Maine Coon-ish characteristics (ringed). Of the 90 non-Maine Coons (gray), 18 also have those characteristics (ringed). So of the 27 ringed cats, only 9—one third—are actually Maine Coons.

The moral is: don't conclude that a hypothesis H is true just because you have some evidence E which is predicted very well by the hypothesis—i.e., $P(E | H)$ is very high. How strongly E favors H depends not on $P(E | H)$ alone but on how it compares with $P(E | \neg H)$, that is, on the *likelihood ratio* $\frac{P(E|H)}{P(E|\neg H)}$. (In Leona's case the ratio is $\frac{9/10}{1/5} = \frac{9}{2}$: her features are four and a half times as likely if she is a Maine Coon as if she is not.) The likelihood ratio is *one* of the factors that, according to Bayes' theorem, determine the posterior— $P(H | E)$. But the other important factor is the prior $P(H)$. This prior is also sometimes called the "base rate", and concluding that $P(H | E)$ must be high just because $P(E | H)$ is high is sometimes called the *fallacy of base rate neglect*.

6.3 Eikosograms

So far we have reasoned about probabilities entirely algebraically—writing down axioms, deriving theorems, and grinding through fractions like the ones in Leona’s case. That is one way to get the right answer, but it is easy to lose the forest for the trees: it is not always obvious *why* $P(M \mid L) = 1/3$ feels surprising, or why it should not. In this section we introduce a picture—the *eikosogram*—that lets us see probabilities as areas, and that makes facts like Leona’s fairly small posterior almost obvious to the eye.

An eikosogram is a square of area 1 that allows us to compare the probabilities of each possible world determined by two propositions φ and ψ . The square is divided into a left strip and a right strip by a vertical line. We then divide each of the strips into a top and bottom part. The lines dividing the left and right strip need not match up with each other. The four resulting rectangles—bottom-left, top-left, bottom-right, and top-right—correspond to $\varphi \wedge \psi$, $\varphi \wedge \neg\psi$, $\neg\varphi \wedge \psi$, and $\neg\varphi \wedge \neg\psi$, respectively. The probability of each of these four molecular propositions (which, you’ll notice, are the rows of the corresponding truth-tables) is the area of the corresponding rectangle. The width of the left strip corresponds to $P(\varphi)$ and that of the right strip to $P(\neg\varphi)$. $P(\psi)$ corresponds to the (possibly irregular) shape that is the combination of the two bottom rectangles. The height of the bottom-left rectangle is $P(\psi \mid \varphi)$, and the height of the bottom-right rectangle is $P(\psi \mid \neg\varphi)$.

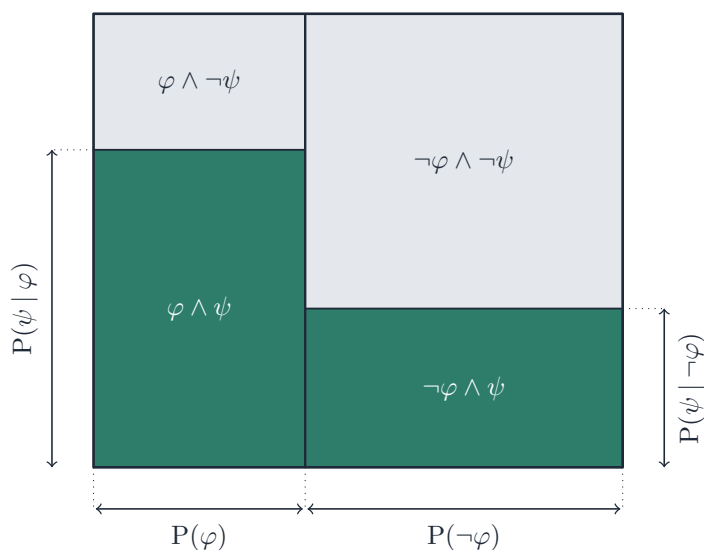


Figure 6.2: An eikosogram for φ and ψ . The shaded region—the two bottom rectangles taken together—corresponds to $P(\psi)$.

That is why eikosograms are useful: every probability we might want—an unconditional probability like $P(\varphi)$, a conjunction like $P(\varphi \wedge \psi)$, even a conditional probability like $P(\psi \mid \varphi)$, which is just the height of a strip—can be read directly off the diagram, as a length or an area.

Let us return to Leona. Recall that $P(M) = 1/10$, $P(L \mid M) = 9/10$, and $P(L \mid \neg M) = 2/10$. Figure 6.3 shows the resulting eikosogram.

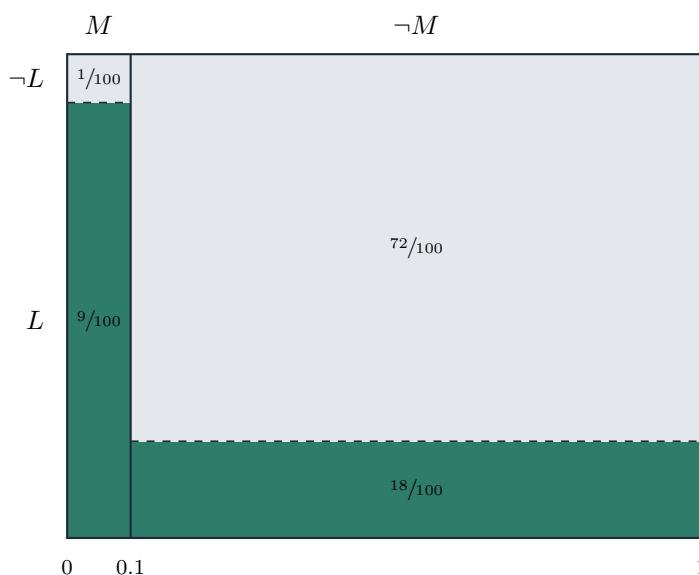


Figure 6.3: The eikosogram for M (is a Maine Coon) and L (has Maine Coon-ish characteristics). The narrow left strip is M ; the tall dark rectangle within it is $M \wedge L$. The wide right strip is $\neg M$; its shorter dark rectangle is $\neg M \wedge L$.

The picture makes both parts of our earlier result vivid at once. First, the dark (“has characteristics”) rectangle in the M strip takes up nearly all of that strip’s height—because $P(L \mid M) = 9/10$ is large, having the characteristics really is good evidence of being a Maine Coon. Second, the M strip itself is thin: because $P(M) = 1/10$ is small, even a tall dark rectangle within it covers only a small area, $9/100$. Meanwhile the dark rectangle in the much wider $\neg M$ strip, though comparatively short, covers a bigger area, $18/100$, simply because it has so much more width to work with. Geometrically, Bayes’ theorem is nothing more than comparing one dark rectangle to the sum of both.

6.4 Hume on miracles

In a very famous passage of his *Enquiry Concerning Human Understanding*, David Hume writes:

No testimony is sufficient to establish a miracle, unless the testimony be of such a kind, that its falsehood would be more miraculous, than the fact, which it endeavours to establish. (...) When any one tells me, that he saw a dead man restored to life, I immediately consider with myself,

whether it be more probable, that this person should either deceive or be deceived, or that the fact, which he relates, should really have happened. I weigh the one miracle against the other; and according to the superiority, which I discover, I pronounce my decision, and always reject the greater miracle. If the falsehood of his testimony would be more miraculous, than the event which he relates; then, and not till then, can he pretend to command my belief or opinion.

In this section we will analyze this passage using what we've learned in this part.

A miracle, for Hume, is a violation of a law of nature. And Hume's conception of a law of nature is a regularity that has been found by unexceptional experience. For instance, a man's being resurrected would be a miracle. It's important to note that Hume does not think that miracles are *impossible*. A miracle is something that goes against all of our vast experience so far, but is not a contradiction.

Hume's argument seems to be the following:

- (1) If testimony is sufficient to establish a miracle, then it's less likely that the testimony is false than that the miracle occurred.
- (2) It is always more likely that the testimony is false than that a miracle occurred.

$\therefore$ No testimony is sufficient to establish a miracle.

Let us start our analysis with a dictionary. Which atomic propositions are used in Hume's argument? It helps to start with the conclusion: that no testimony is sufficient to establish a miracle. This is clearly the negation that testimony is (sometimes) sufficient to establish a miracle. How could we symbolize in **PL** that proposition? One possibility is as a conditional: if someone testifies that a miracle occurred, then a miracle occurred. The first premise then says that if a miracle can ever be established by testimony, then it's less likely that the testimony is false than that the miracle occurred. Our dictionary, then, can be the following:

T = Someone testifies that a miracle occurred.

M = A miracle occurred.

F = It's less likely that the testimony is false than that a miracle occurred.

We can then reconstruct the argument as follows:

$$(1) (T \rightarrow M) \rightarrow F$$

$$(2) \neg F$$

$$\therefore \neg M$$

The argument is valid, as you can verify by constructing its truth-table. Now, what about the premises? Let's start with the first one:

If testimony is sufficient to establish a miracle, then it's less likely that the testimony is false than that the miracle occurred.

What does it mean to say that testimony is sufficient to establish a miracle? We can assume that if our credence that a miracle has happened is less than 50%, then the miracle is not established, so let us say that testimony T establishes a miracle M (for us) just in case our degree of belief in M given T is at least 50%:

$$\text{A miracle has been established} = \text{Cr}_{old}(M | T) \geq 0.5$$

Then we receive as evidence the testimony T that the miracle in question has occurred. What should our degree of belief in M be now? If we are Bayesians, it should be:

$$\text{Cr}_{new}(M) = \text{Cr}_{old}(M | T) = \frac{\text{Cr}_{old}(T | M) \times \text{Cr}_{old}(M)}{\text{Cr}_{old}(T)} \quad (1)$$

In the same way, our credence that no miracle has occurred after hearing testimony that it has is given by:

$$\text{Cr}_{new}(\neg M) = \text{Cr}_{old}(\neg M | T) = \frac{\text{Cr}_{old}(T | \neg M) \times \text{Cr}_{old}(\neg M)}{\text{Cr}_{old}(T)} \quad (2)$$

Dividing (1) by (2), the unknown $\text{Cr}_{old}(T)$ cancels out, which means that the odds of our credence that a miracle has occurred to our credence that it has not (after hearing testimony in its favor) are given by the following:³

$$\frac{\text{Cr}_{old}(M | T)}{\text{Cr}_{old}(\neg M | T)} = \frac{\text{Cr}_{old}(T | M)}{\text{Cr}_{old}(T | \neg M)} \times \frac{\text{Cr}_{old}(M)}{\text{Cr}_{old}(\neg M)} \quad (3)$$

³Throughout we assume that none of the quantities we divide by is 0—that is, that $\text{Cr}_{old}(T)$, $\text{Cr}_{old}(\neg M)$, $\text{Cr}_{old}(T | M)$, and $\text{Cr}_{old}(T | \neg M)$ are all greater than 0.

Since Cr_{new} is a probability function, Theorem 5.5 gives $\text{Cr}_{\text{new}}(\neg M) = 1 - \text{Cr}_{\text{new}}(M)$, and so $\text{Cr}_{\text{old}}(\neg M | T) = 1 - \text{Cr}_{\text{old}}(M | T)$:

$$\text{Cr}_{\text{old}}(M | T) \geq 0.5 \quad \text{iff} \quad \frac{\text{Cr}_{\text{old}}(M | T)}{\text{Cr}_{\text{old}}(\neg M | T)} \geq 1 \quad (4)$$

Now, from (3) and (4) it follows that:

$$\text{Cr}_{\text{old}}(M | T) \geq 0.5 \quad \text{iff} \quad \frac{\text{Cr}_{\text{old}}(T | M)}{\text{Cr}_{\text{old}}(T | \neg M)} \times \frac{\text{Cr}_{\text{old}}(M)}{\text{Cr}_{\text{old}}(\neg M)} \geq 1 \quad (5)$$

Rearranging (5):

$$\text{Cr}_{\text{old}}(M | T) \geq 0.5 \quad \text{iff} \quad \frac{\text{Cr}_{\text{old}}(M)}{\text{Cr}_{\text{old}}(\neg M)} \geq \frac{\text{Cr}_{\text{old}}(T | \neg M)}{\text{Cr}_{\text{old}}(T | M)} \quad (6)$$

Equivalently:

$$\text{Cr}_{\text{old}}(M | T) \geq 0.5 \quad \text{iff} \quad \text{Cr}_{\text{old}}(T | \neg M) \leq \text{Cr}_{\text{old}}(T | M) \cdot \frac{\text{Cr}_{\text{old}}(M)}{\text{Cr}_{\text{old}}(\neg M)} \quad (7)$$

If we assume that the testifier will definitely testify to the miracle if he experiences it (i.e., that $\text{Cr}_{\text{old}}(T | M) \approx 1$ —notice that this assumption is friendly to the believer in miracles), we have that:

$$\text{Cr}_{\text{old}}(M | T) \geq 0.5 \quad \text{iff} \quad \text{Cr}_{\text{old}}(T | \neg M) \lesssim \frac{\text{Cr}_{\text{old}}(M)}{\text{Cr}_{\text{old}}(\neg M)} \quad (8)$$

If we assume with Hume that $\text{Cr}_{\text{old}}(M) \approx 0$ (but positive!), then $\frac{\text{Cr}_{\text{old}}(M)}{\text{Cr}_{\text{old}}(\neg M)} \approx \text{Cr}_{\text{old}}(M)$, and so:

$$\text{Cr}_{\text{old}}(M | T) \geq 0.5 \quad \text{iff} \quad \text{Cr}_{\text{old}}(T | \neg M) \lesssim \text{Cr}_{\text{old}}(M) \quad (9)$$

And (9) can be seen as a probabilistic reconstruction of Hume's premise 1: a testimony can establish a miracle just in case its falsity is less likely than the miracle in question. Call (9) *Hume's test*. So the test itself is (a strengthening of) premise 1, and Hume's contention that no testimony passes that test (i.e., that the falsity of a testimony is always more likely than the occurrence of a miracle) is his premise 2. So, is his premise 2 true? Let us try to put some numbers into the probabilities.

Let us suppose that miracles are 1 in 10,000—a very liberal view of miracles that is not at all friendly to Hume. In that case, $\text{Cr}_{\text{old}}(M) = \frac{1}{10,001}$, and, correspondingly, $\text{Cr}_{\text{old}}(\neg M) = \frac{10,000}{10,001}$. Let us also assume that if a miracle did in fact occur, then it will for sure be testified to: $\text{Cr}_{\text{old}}(T | M) = 1$. Hume's crucial point is that,

nevertheless, there will be some false testimonies—due to intentional misleadingness, or self-deception, or whatever. So let us suppose that testimony to a miracle is rarely produced when no miracle has occurred: in 1 out of every 100 cases in which there is no miracle, there is nonetheless testimony that there was one: $\text{Cr}_{old}(T \mid \neg M) = \frac{1}{100}$.

Now we compute the posterior using the LTP version of Bayes' theorem:

$$\begin{aligned}
 \text{Cr}_{new}(M) = \text{Cr}_{old}(M \mid T) &= \frac{\text{Cr}_{old}(T \mid M) \times \text{Cr}_{old}(M)}{\text{Cr}_{old}(T \mid M) \times \text{Cr}_{old}(M) + \text{Cr}_{old}(T \mid \neg M) \times \text{Cr}_{old}(\neg M)} \\
 &= \frac{1 \times \frac{1}{10,001}}{1 \times \frac{1}{10,001} + \frac{1}{100} \times \frac{10,000}{10,001}} \\
 &= \frac{\frac{1}{10,001}}{\frac{1}{10,001} + \frac{100}{10,001}} \\
 &= \frac{1}{101} \\
 &\approx 0.00990099 (= 0.990099\%).
 \end{aligned}$$

So, even with assumptions that are friendly to the existence of miracles and the veracity of testifiers, Hume's point seems to stand: our credence in a miracle on the basis of testimony for it should be less than 0.01. The eikosogram would look something like this (although we cannot really represent how thin the left rectangle would have to be and keep it visible):

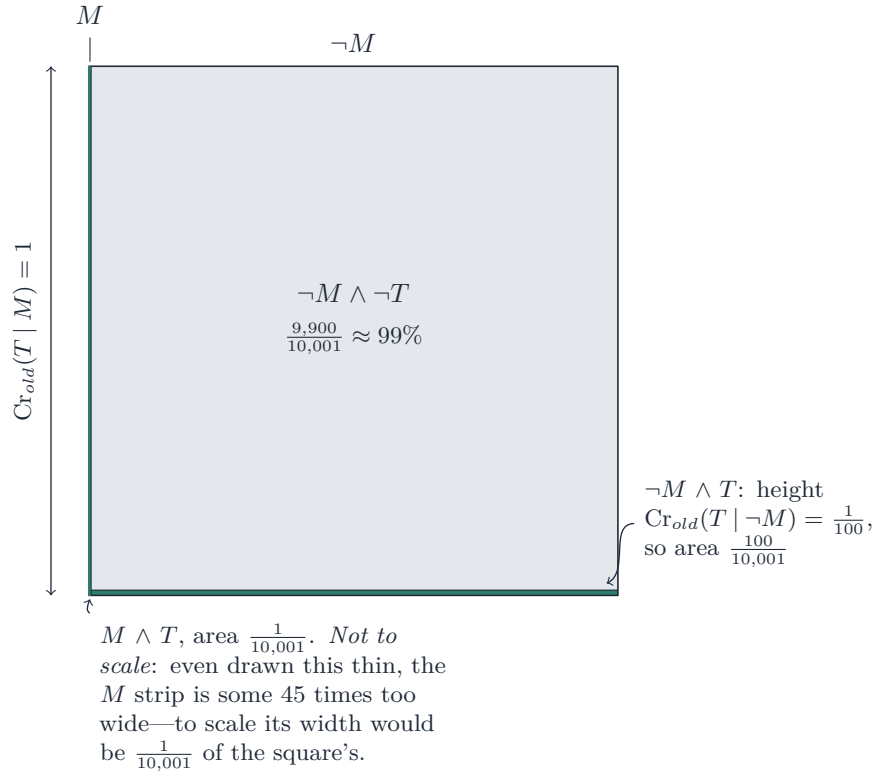


Figure 6.4: The eikosogram for M (a miracle occurs) and T (a miracle is testified to). The shaded region is T . Because $\text{Cr}_{old}(T | M) = 1$, the M strip is shaded to its full height, and the rectangle for $M \wedge \neg T$ has area 0. The shaded region is overwhelmingly made up of false testimony: of its total area $\frac{101}{10,001}$, only $\frac{1}{10,001}$ lies in the M strip—which is why $\text{Cr}_{old}(M | T) = \frac{1}{101}$.

But what if what we receive as evidence is not just a single testimony, but several? A crucial point here is whether these testifiers give us *independent* evidence or not. If a second testifier is just repeating what the first one says because she believes him, that should not add any evidence in favor of the miracle. Let us suppose, then, that we receive two testimonies T_1 and T_2 , and that the testifiers are conditionally independent given M and given $\neg M$:

$$\begin{aligned}\text{Cr}_{old}(T_1 \wedge T_2 | M) &= \text{Cr}_{old}(T_1 | M) \times \text{Cr}_{old}(T_2 | M), \\ \text{Cr}_{old}(T_1 \wedge T_2 | \neg M) &= \text{Cr}_{old}(T_1 | \neg M) \times \text{Cr}_{old}(T_2 | \neg M).\end{aligned}$$

Assume, as before, that $\text{Cr}_{old}(T_1 | M) = \text{Cr}_{old}(T_2 | M) = 1$ and that $\text{Cr}_{old}(T_1 | \neg M) =$

$\text{Cr}_{old}(T_2 \mid \neg M) = \frac{1}{100}$. Then:

$$\begin{aligned}\text{Cr}_{old}(T_1 \wedge T_2 \mid M) &= 1 \times 1 = 1, \\ \text{Cr}_{old}(T_1 \wedge T_2 \mid \neg M) &= \frac{1}{100} \times \frac{1}{100} = \frac{1}{10,000}.\end{aligned}$$

By the LTP version of Bayes' theorem:

$$\text{Cr}_{old}(M \mid T_1 \wedge T_2) = \frac{\text{Cr}_{old}(T_1 \wedge T_2 \mid M) \text{Cr}_{old}(M)}{\text{Cr}_{old}(T_1 \wedge T_2 \mid M) \text{Cr}_{old}(M) + \text{Cr}_{old}(T_1 \wedge T_2 \mid \neg M) \text{Cr}_{old}(\neg M)}.$$

Substituting:

$$\begin{aligned}\text{Cr}_{new}(M) = \text{Cr}_{old}(M \mid T_1 \wedge T_2) &= \frac{1 \times \frac{1}{10,001}}{1 \times \frac{1}{10,001} + \frac{1}{10,000} \times \frac{10,000}{10,001}} \\ &= \frac{\frac{1}{10,001}}{\frac{1}{10,001} + \frac{1}{10,001}} \\ &= \frac{1}{2} = 0.5.\end{aligned}$$

Notice that this lands exactly on the threshold, and not by accident: $\text{Cr}_{old}(T_1 \wedge T_2 \mid \neg M) = \frac{1}{10,000}$, which is precisely $\frac{\text{Cr}_{old}(M)}{\text{Cr}_{old}(\neg M)}$. The pair of testimonies passes Hume's test with nothing to spare. A third independent testimony of the same quality would give $\text{Cr}_{new}(M) = \frac{100}{101} \approx 0.99$.

These results obviously depend on the numbers we plug in, but the pattern is clear: independent testimonies can overwhelm even a very small prior. Whether that helps the believer in miracles is another question—for it is precisely the independence of testimonies that is hardest to come by.

6.5 The problem of priors

As we saw, which prior credences we have greatly affects the results of applying Bayes' theorem—and, thus, the results of abiding by the diachronic norm for credences. For instance, if your priors were that Maine Coons make up 50% of the feline population, you would have concluded that it was 82% likely that Leona was a Maine Coon given that she has Maine Coon-ish features. So, the crucial question is: where do the priors come from?

To appreciate the importance of this question, we must first dismiss the temptation to think that it has an obvious answer. The temptation is to say that priors come either from past experience, or past evidence or information that we have.

For instance, in the case of Leona, I told you that Maine Coons make up about 10% of the feline population. You believed me, and that is where your priors came from. Where is the mystery?

There is perhaps no psychological mystery, in that it may well have been that, as a matter of fact, you did form your priors in the Leona case just by trusting what I told you. But the problem of priors is *normative*, not psychological. That is to say: we are not worried about the psychological process by which subjects come to form their prior credences; rather, we are worried about how they *rationally should* form those prior credences. Why did you believe me when I told you about the base rate of Maine Coons? Well, the Bayesian answer must be that you had a prior for *that* as well: you thought that the conditional probability of the base rate being x , given that I said that it was x , was very high—perhaps even 1. OK, but why did you have *that* prior? Perhaps because you had yet an earlier prior about testimony from other humans in general. But why did you have that one? Pretty soon, we are asking about the most basic priors one can have: the conditional probabilities that dictate how someone without any empirical information whatsoever will react to any possible piece of empirical information. Sometimes, these are called *ur-priors*. So the problem of priors is really the problem of ur-priors: what determines what are the fundamental ways we react to evidence?

There are broadly two answers to this question. *Subjective Bayesians* think that *any* prior will do, as long as it satisfies the probability calculus. According to subjective Bayesians, for instance, it doesn't matter whether you think that all cats are Maine Coons, that none are, or anything in between.⁴ Doesn't this inject an absurd measure of arbitrariness in our credences? Subjective Bayesians respond by appealing to *convergence theorems*, according to which it doesn't matter what our ur-priors are because even two subjects who start out with very different ur-priors will converge in their posteriors after exposure to sufficient amounts of evidence.

It's not obvious that this appeal to convergence theorems can save subjective Bayesianism from the charge of arbitrariness, for several related reasons. First, the subjects cannot start with *any* priors in order to guarantee convergence—they must agree on which possibilities to rule out altogether: neither can assign probability zero to some possibility that the other still takes seriously, and vice versa. That is already a severe restriction on priors, and it is not clear why it fits with the subjective Bayesian picture. Second, and this is a different sort of

⁴Some subjective Bayesians do add one caveat: your ur-priors should be *regular*: they should not assign extreme probabilities (1 or 0) to any contingent proposition.

worry than the first: even granting that two subjects' priors satisfy that restriction, so that convergence is guaranteed to occur between them, it does not follow that convergence covers every proposition they have opinions about. The convergence relates only to those propositions whose credence can be affected by evidence. If a hypothesis for a subject is completely independent of any possible evidential stream, then it will retain that prior no matter what. Thus, two subjects with different priors over hypotheses that are independent in this way will never converge on the credence of these hypotheses. Moreover, it may be that a hypothesis is independent of the evidence stream for one subject, but not for another. These two subjects will also not converge on the credence for this hypothesis. Third, and finally, the convergence theorems are limit theorems. What this means is that credences tend to converge as the number of propositions acquired as evidence tends to infinity. This guarantees absolutely nothing about what happens before the long run. And, as Keynes famously said, "in the long run we are all dead."

So it is not obvious that Subjective Bayesians can satisfactorily answer the problem of priors. Rejecting Subjective Bayesianism commits one only to thinking that there are *some* restrictions on ur-priors. So there is a spectrum with Subjective Bayesianism on one end and what we will call *Objective Bayesianism* on the other: the view according to which there is only one legitimate ur-prior which determines how every rational subject should respond to every possible stream of evidence. If the problem for Subjective Bayesianism was the arbitrariness it induces, the problem for Objective Bayesianism is the opposite: it makes it impossible for rational agents to "agree to disagree" even on the most minute matter of evidential import. This raises an epistemological problem and a related metaphysical one. The epistemological problem is: how can we figure out what the unique rational ur-prior is? And the metaphysical problem is: what could even settle what that unique rational prior is?

One answer to both the epistemological and the metaphysical question with some historical precedent is related to the "Principle of Indifference". Applied to our framework, this principle says that, in the absence of any relevant evidence, we should assign the same probability to every row of the truth-table. There are several serious problems with the Principle of Indifference. One is that it makes every atomic proposition independent of every other one and that means that receiving an atomic proposition as evidence can never make us change our minds about any other atomic proposition. Objective Bayesians can try to get around this problem by saying that equal probabilities should be distributed not over every

row of the truth-table, but over some other partition of the possibilities. In any case, however, the Principle of Indifference must face another serious objection.

Consider a factory that produces cubical boxes that range in height from 1 to 2 ft. tall. If you had to predict the height of the next box, what would you say? An Objective Bayesian would advise you to apply the Principle of Indifference and give the same probability to each possible height. Therefore, you should predict that the next box will be 1.5 ft. tall. But if the boxes range in height from 1 to 2 ft. tall, they also range in volume from 1 ft.³ to 8 ft.³ (given that the volume of a 2 ft. tall cube is $2^3 = 8$). If you assign equal probabilities to every possible volume the next cube might have, you will predict the next cube to have 4.5 ft.³. But this answer conflicts with the previous one: the volume of a 1.5 ft. tall box is 3.375 ft.³, not 4.5 ft.³.

The problem of the priors, then, doesn't seem to have an easy solution. There are sophisticated versions of both Subjective and Objective Bayesianism that attempt answers to the problems canvassed here—but for those, you would have to take another class.

Exercises

For each of the following, start by giving a dictionary, then identify the priors and the likelihoods you have been given, and then use Bayes' theorem to compute the posterior.

A common symptom of bubonic plague is the appearance of “buboes”—lymph nodes swollen to the size of eggs on one’s neck and elsewhere. 70% of individuals who have the plague also present with buboes, while 10% of plague-free individuals in the area of an outbreak experience buboe-like swellings because of their limited exposure to the disease. Due to a recent outbreak, about 1 in every 20 people in the area have contracted bubonic plague.

1. If you wake up one morning with an egg-sized swelling on your neck, how much credence should you give to the possibility that you have the bubonic plague?
2. Suppose that no plague-free individuals experience buboe-like swellings. With everything else being the same as in the original story, how much credence

should you give to the possibility that you have the bubonic plague given that there is an egg-sized swelling on your neck?

3. Suppose that you don't have any swelling on your neck. With everything else being the same as in the original story, how much credence should you give to the possibility that you have the bubonic plague given that you have no swelling?

A police officer pulls over a speeding car late at night on Thanksgiving. Historically, 30% of drivers who are caught speeding on Thanksgiving are also drunk. The officer asks the driver to recite the alphabet backwards, and the driver fails to do so. No drunk person can recite the alphabet backwards, and only 80% of sober people can reliably recite the alphabet backwards.

4. Considering that they were caught speeding on Thanksgiving, how confident should the officer be that the driver is also drunk given that they couldn't recite the alphabet backwards?
5. Suppose that 5% of drunk people can recite the alphabet backwards and 90% of sober people can do it. With everything else being the same as in the original story, how confident should the officer be that the driver is drunk given that they couldn't recite the alphabet backwards?
6. Suppose that historically only 20% of drivers who are caught speeding on Thanksgiving are drunk. With everything else being the same as in the original story, how confident should the officer be that the driver is drunk given that they couldn't recite the alphabet backwards?

There was a hit and run involving a taxicab, and a witness reported that the cab was from Citi Taxi, which operates a third of the cabs in the city. Witnesses in the city are reliable only 80% of the time.

7. Given the witness testimony, how confident should we be that a Citi Taxi cab was involved in the incident?
8. Suppose that Citi Taxi cabs are very distinctive, so that witnesses in the city will correctly identify a Citi Taxi cab 80% of the time, and only misidentify a cab as belonging to Citi Taxi 5% of the time. With everything else being the

same as in the original story, how confident should we be that a Citi Taxi cab was involved in the incident given the witness testimony?

According to the Intentional Theory of Art, something is art if and only if its creator intended for it to be regarded as art. Some practice sketches are presented in museums, but it's not clear if they were ever intended to be regarded as art. A recent poll of major artists has revealed that only 20% of them intend for their practice sketches to be regarded as art. Among those who intend for their practice sketches to be regarded as art, only 5% of their practice sketches have appeared in museums. Similarly, museums have featured 5% of practice sketches by people who don't intend for their sketches to be regarded as art.

9. How confident should we be that a practice sketch is actually art according to the Intentional Theory given that it is on display in a museum?
10. According to the Institutional Theory of Art, something is art if and only if it is declared to be art by a member of the "artworld". Displaying something in a museum is a standard way for members of the artworld to declare that the thing being displayed is art. With everything else being the same as in the original story, how confident should we be that a practice sketch is actually art *according to the Institutional Theory* given that it is on display in a museum?

Part III

Decision and Game Theory

7 Decisions

This chapter sets out the mathematical model of a decision (choices, states, outcomes, and utilities), represents decisions with matrices, and shows how ordinal utilities can be read off an agent's preferences when those preferences form a weak order. It then turns to choice under ignorance, where we cannot assign probabilities to the states, and compares three norms: dominance, maximin (with its refinement, leximin), and minimax regret.

In Part I we learned about propositional logic and how it helps us analyze arguments and determine their tautological validity. In Part II we learned about probability, and how it helps us analyze probabilistic arguments. In this part, we'll apply everything we learned in the previous two, plus some new techniques, to analyze how to make rational decisions. We start by considering those decisions that involve only one agent as a decision-maker, and then we move on to interactive decision theory—also known as game theory.

We start by setting out the main elements in our mathematical model of decisions. They are: *choices*, *states*, *outcomes*, and *utilities*. *Choices* are the options that the agent has: go to college, work on the family farm, or form a rock band, for instance. *States* are ways the world can be that affect what outcome your choice will have: whether the economy will tank in the next five years or not, for example. *Outcomes* are, precisely, a function of your choice on the one hand and the state of the world on the other: for instance, if you decide to work on the family farm and the economy tanks, the outcome is not as good as if it did not. Finally, *utilities* are a measure of your preferences over outcomes.

7.1 Representing decisions

The standard way of representing decisions is by using a *matrix*. For a decision with n choices and m states, we use an n by m matrix where the choices label the rows, the states label the columns and, the cells represent outcomes:

		states			
		S_1	S_2	S_3	S_4
choices	A_1	O_{11}	O_{12}	O_{13}	O_{14}
	A_2	O_{21}	O_{22}	O_{23}	O_{24}
	A_3	O_{31}	O_{32}	O_{33}	O_{34}

↑
outcomes

Table 7.1: A decision problem with three choices and four states. Each cell records the outcome of making the choice on its row in the state in its column.

In general, we will write the utilities associated with the outcomes in each cell instead:

		states			
		S_1	S_2	S_3	S_4
choices	A_1	$u(O_{11})$	$u(O_{12})$	$u(O_{13})$	$u(O_{14})$
	A_2	$u(O_{21})$	$u(O_{22})$	$u(O_{23})$	$u(O_{24})$
	A_3	$u(O_{31})$	$u(O_{32})$	$u(O_{33})$	$u(O_{34})$

Table 7.2: The same decision problem, with each cell now recording the utility of the outcome that cell represents.

Suppose that we have a choice between going to the park or going to the movies. Those are our *choices*. It could rain or not. Those are the *states*. Pairing a choice with a state gives us an *outcome*, and so there are four of them: going to the park while it rains, going to the park while it doesn't rain, going to the movies while it rains, and going to the movies while it doesn't rain. Let us assume that your utilities for those four outcomes are 1, 4, 3, and 2 respectively (we will study where those utilities come from later in this chapter). A rained-out afternoon at the park is the worst thing that could happen to you, and a sunny one is the best; the movies are somewhere in the middle either way, and marginally better when it rains, since then at least you are not squandering good weather indoors. Writing P for going to the park, M for going to the movies, and R for its raining, we can represent this decision problem with the following matrix:

	R	$\neg R$
P	1	4
M	3	2

Table 7.3: The decision problem of whether to go to the park (P) or to the movies (M), depending on whether it rains (R) or not. The rows are the choices, the columns are the states, and each cell records the utility of the corresponding outcome.

Not just any list of propositions will do for the columns of such a matrix. Suppose we had taken our states to be *it rains* and *it snows*. Then a clear day is a way things might turn out that no column covers, and the matrix simply says nothing about what would happen if it came about—our list of outcomes would be incomplete. Suppose instead that we had taken our states to be *it rains* and *it is cold*. Now a cold, rainy day falls under both columns at once, and the cells disagree with each other about what its outcome is. To rule out both defects, we require that the states *partition* the possibility space:

Definition 7.1. Partition: A set of propositions $\{S_1, S_2, \dots, S_n\}$ is a *partition* of the possibility space if and only if exactly one of them is true in each possible world. That is to say, they must be both:

1. *Mutually exclusive:* no two of them are true in the same possible world (see Definition 4.1); and
2. *Jointly exhaustive:* there is no possible world in which none of them is true.

Mutual exclusivity is what rules out the second bad case, and joint exhaustiveness what rules out the first. Together they guarantee that, however things turn out, exactly one column describes the situation you are actually in—which is what entitles us to read each row as a complete specification of what a choice would get you.

Our original example passes the test, and does so for a reason that will often be useful: a proposition and its negation always form a partition, since as a matter of logic R and $\neg R$ cannot both be true and cannot both be false. The rows of a truth-table are a partition as well: exactly one row describes any given possible world, which is precisely why the probabilities we assign to the rows of a stochastic truth-table have to sum to 1.

7.2 Where do ordinal utilities come from?

Where did the utility numbers come from? The answer is: from your preferences among outcomes. Utilities are a scale we use to represent your preferences. We will consider two different kinds of utility scales, *ordinal* and *cardinal*. We start with ordinal scales in this chapter, and turn to cardinal scales in Chapter 8.

We will use the symbol " $\succ$ " to represent preference relations— $\varphi \succ \psi$ means that the agent prefers φ to ψ . If an agent is indifferent between φ and ψ , we will write $\varphi \sim \psi$. Recall that your most preferred option is a sunny day at the park, which you prefer even to the movies when it rains; that you prefer being at the movies when it rains to being at the movies when it doesn't; and that you prefer being at the movies when it doesn't rain even to being at the park when it rains. We can therefore represent these preferences as follows:

$$\begin{aligned}(P \wedge \neg R) &\succ (M \wedge R) \\ (M \wedge R) &\succ (M \wedge \neg R) \\ (M \wedge \neg R) &\succ (P \wedge R)\end{aligned}$$

Notice that we have not stated all of your preferences here, because we have not stated that you prefer being at the park when it doesn't rain to being at the movies when it doesn't rain. You might think that goes without saying, given that you prefer being at the park when it doesn't rain to being at the movies when it rains, and you prefer being at the movies when it rains to being at the movies when it doesn't rain. That is to say, you might think it goes without saying that your preferences should be *transitive*. If so, decision theorists agree with you. In fact, decision theory presupposes that your preferences satisfy several global constraints:

Definition 7.2. Asymmetric: An agent's preferences are asymmetric if and only if, for any two outcomes φ and ψ :

1. If $\varphi \succ \psi$, then not $\psi \succ \varphi$.
2. If $\varphi \succ \psi$, then not $\psi \sim \varphi$.
3. If $\varphi \sim \psi$, then not $\varphi \succ \psi$ and also not $\psi \succ \varphi$.

Definition 7.3. Transitive: An agent's preferences are transitive if and only if, for any three outcomes χ , φ and ψ :

1. If $\varphi \succ \psi$ and $\psi \succ \chi$, then $\varphi \succ \chi$.
2. If $\varphi \sim \psi$ and $\psi \sim \chi$, then $\varphi \sim \chi$.
3. If $\varphi \succ \psi$ and $\psi \sim \chi$, then $\varphi \succ \chi$.
4. If $\varphi \sim \psi$ and $\psi \succ \chi$, then $\varphi \succ \chi$.

Definition 7.4. Connected: An agent's preferences are connected if and only if, for any two outcomes φ and ψ , at least one of the following holds:

1. $\varphi \succ \psi$.
2. $\psi \succ \varphi$.
3. $\varphi \sim \psi$.

Notice that Definition 7.2 and Definition 7.4 work together: connectedness guarantees that at least one of the three relations holds between any two outcomes, and asymmetry guarantees that at most one of them does. Taken together, then, they tell us that for any two outcomes exactly one of $\varphi \succ \psi$, $\psi \succ \varphi$, and $\varphi \sim \psi$ is true—an agent whose preferences satisfy both is never undecided between two outcomes, and never prefers each of two outcomes to the other.

Notice also the crucial distinction between being indifferent between two outcomes, on the one hand, and having no preference relations whatsoever between them, not even indifference. If you are indifferent between two outcomes, that means that you are completely willing to replace one with the other. Right now, for instance, I am indifferent between having coffee and tea—I could go for either one. On the other hand, some people have no preference relations between The Rolling Stones and Mozart. It's not that they are indifferent between them—it's rather that they think that they each have different properties that make them *incommensurable*—they just cannot compare one to the other. For better or worse, Definition 7.4 prohibits these incommensurable outcomes. Normal humans also sometimes have intransitive preferences. Suppose you add a single grain of sugar to successive cups of coffee. Normal humans will be indifferent between neighboring cups of coffee, but will definitely have a preference between the first and the last.

What does the fact that we violate some of these conditions mean for the theory of decision that presupposes that we don't? We won't deal with this issue here at all, except to make the following two points: first, decision theory is at least partly *normative*—it tells us not how our preferences actually are ordered, but how

they *should* be ordered if we are to make rational decisions; second, there is some comfort in the fact that we can make rational decisions even when we don't satisfy all the structural conditions on preferences—for instance, even if your preferences are not connected among all the possible choices for dessert, but you know you can have your first choice, the lack of connectedness doesn't matter for that particular choice.

If an agent's preferences do satisfy all the structural conditions on preferences, we will say that his preferences form a *weak order*:

Definition 7.5. Weak order: An agent's preferences form a weak order if and only if they are connected, asymmetric, and transitive.

If your preferences form a weak order, then they can be represented by a ranking. The agent's preferences in movies/park case, for example, can be represented in the following ranking (ordered from best to worst top to bottom):

- $P \wedge \neg R$
- $M \wedge R$
- $M \wedge \neg R$
- $P \wedge R$

Of course, the structural definitions allow ties, but in this case the agent is not indifferent between any of the outcomes.

Once we have a ranking (with ties allowed), we can assign a number to each outcome reflecting its place in the ranking, with higher numbers assigned to higher-ranked outcomes. For instance, we could have the following numbers assigned to the outcomes for the movies/park case:

4. $P \wedge \neg R$
3. $M \wedge R$
2. $M \wedge \neg R$
1. $P \wedge R$

Other than assigning higher numbers to higher-ranked outcomes, we are free to assign whatever number we wish. For instance, this other assignment is just as good as the previous one:

2000 $P \wedge \neg R$

1395 $M \wedge R$

2 $M \wedge \neg R$

-12 $P \wedge R$

We call the number assigned to an outcome its *utility*, and we write it $u(O_i) = n$. If the utility function u satisfies the requirement just laid out, we say that it is an *ordinal* utility function:

Definition 7.6. Ordinal utility function: A utility function u is *ordinal* if and only if, for any two outcomes φ and ψ :

- a $u(\varphi) > u(\psi)$ if and only if $\varphi \succ \psi$; and
- b $u(\varphi) = u(\psi)$ if and only if $\varphi \sim \psi$.

The numbers in an ordinal utility function represent nothing more and nothing less than the order of the corresponding outcome in the ranking of the agent's preferences. If u is a correct ordinal representation of an agent's preferences, then any *ordinal transformation* u' of u will also be a correct ordinal representation of those preferences:

Definition 7.7. Ordinal transformation: u' is an ordinal transformation of u if and only if, for every two outcomes φ and ψ : $u'(\varphi) \geq u'(\psi)$ if and only if $u(\varphi) \geq u(\psi)$

7.3 Choice under ignorance

Decision theorists usually distinguish between three different kinds of decisions: decisions under certainty, decisions under ignorance, and decisions under risk. A decision counts as a *decision under certainty* when the agent is certain which state in the partition obtains. Decision under certainty need not be computationally easy (the mathematical theory of linear programming has been fruitfully applied to this topic), but theoretically does not give rise to very interesting topics, and therefore we shall not dwell on it any more.

When agents cannot eliminate all but one of the states in the relevant partition, but can assign probabilities to those states, we are dealing with *decisions under*

risk. We will deal with those kinds of decisions in the next chapter. Finally, when subjects cannot even assign meaningful probabilities to the states of the relevant partition, we are dealing with decisions under ignorance. That is what we will be dealing with in the rest of this chapter.

7.3.1 Dominance

Suppose that, in the movies/park example, we have no idea whether it will rain or not. Having no idea is very different from being 50/50 on whether it will rain or not—the latter would be a decision under risk, whereas if we genuinely have no idea then we are facing a decision under ignorance. You might think that there is not much that can be said about how to go about making decisions when we are completely ignorant regarding the probabilities of the different possible states, but you’d be wrong. Consider, for instance, a slight variation of our movies/park example, one where your ranking (together with a possible ordinal utility function) is instead as follows:

4. $P \wedge \neg R$
3. $P \wedge R$
2. $M \wedge \neg R$
1. $M \wedge R$

The matrix corresponding to that ranking is then the following:

	R	$\neg R$
P	3	4
M	1	2

Table 7.4: Park/movies case with alternate ranking of outcomes.

[Try it in the Decisions Lab ↗](#)

Something remarkable about this ranking of outcomes is that, no matter whether it rains or not, you prefer to go to the park than to go to the movies. So, even if you are completely ignorant of whether it is going to rain or not, you should go to the park: you will then get your preferred outcome no matter what state the world is in.

When that happens—i.e., when one choice results in a better outcome than all the other ones no matter what state obtains—we say that the choice in question *dominates* the other. We will define both a *strong* and a *weak* form of dominance:

Definition 7.8. Strong dominance: Option φ strongly dominates option ψ if and only if, for every state S_i ,

$$u(\varphi \wedge S_i) > u(\psi \wedge S_i).$$

Definition 7.9. Weak dominance: Option φ weakly dominates option ψ if and only if, for every state S_i ,

$$u(\varphi \wedge S_i) \geq u(\psi \wedge S_i),$$

and for at least one state S_j ,

$$u(\varphi \wedge S_j) > u(\psi \wedge S_j).$$

And, on the basis of those definitions, we will propose our first norm for decisions:

Norm 7.10. Dominance norm: dominated options are irrational

Notice that the dominance norm doesn't tell us which option is rational, it only tells us which are irrational—namely, the dominated ones. In the special case where there are only two options, if one of them is dominated then the Dominance norm entails that the other one is the only rationally permissible one. But for cases with more than two options, the fact that one is dominated does not immediately tell you which option is rationally required.

If for two given choices neither one of them is dominated by the other, there are still two other possibilities: it may be that they have the same utility for every state (in which case we'll say that they are equivalent), or it may be that one of them has a higher utility for at least one state, and the other one has a higher utility for at least one other state (in which case we'll say that they are crossing):

Definition 7.11. Equivalent: Options φ and ψ are equivalent if and only if, for every state S_i ,

$$u(\varphi \wedge S_i) = u(\psi \wedge S_i).$$

Definition 7.12. Crossing: Options φ and ψ are crossing if and only if there is some state S_i such that

$$u(\varphi \wedge S_i) > u(\psi \wedge S_i),$$

and some state S_j such that

$$u(\psi \wedge S_j) > u(\varphi \wedge S_j).$$

To see the dominance norm at work, consider the following three decision problems, each involving three choices and three states.

	S_1	S_2	S_3
A_1	1	2	1
A_2	2	3	1
A_3	2	3	2

Table 7.5: A decision problem in which working through the dominance relations pairwise yields a unique solution.

[Try it in the Decisions Lab ↗](#)

We start looking for choices that dominate others by going through them pairwise—that is, we consider each possible pair of choices, ignoring order:

A_1, A_2 : A_2 weakly dominates A_1 : going through each state in turn, $u(A_2 \wedge S_1) > u(A_1 \wedge S_1)$ (since $2 > 1$), $u(A_2 \wedge S_2) > u(A_1 \wedge S_2)$ (since $3 > 2$), and $u(A_2 \wedge S_3) = u(A_1 \wedge S_3)$ (since $1 = 1$).

A_1, A_3 : A_3 strongly dominates A_1 : $u(A_3 \wedge S_1) > u(A_1 \wedge S_1)$ (since $2 > 1$), $u(A_3 \wedge S_2) > u(A_1 \wedge S_2)$ (since $3 > 2$), and $u(A_3 \wedge S_3) > u(A_1 \wedge S_3)$ (since $2 > 1$).

A_2, A_3 : A_3 weakly dominates A_2 : $u(A_3 \wedge S_1) = u(A_2 \wedge S_1)$ (since $2 = 2$), $u(A_3 \wedge S_2) = u(A_2 \wedge S_2)$ (since $3 = 3$), and $u(A_3 \wedge S_3) > u(A_2 \wedge S_3)$ (since $2 > 1$).

The fact that A_2 dominates A_1 does not mean that we should pick A_2 , since A_2 is itself dominated by A_3 . Applying the dominance norm is a matter of ruling out choices until, hopefully, only one is left. Once we see that A_1 is dominated, we can cross it off as something we would never want to do; once we see that A_2 is also dominated, by A_3 , we cross it off too. That leaves us with A_3 as the only remaining choice: dominance reasoning has identified A_3 as the single, or *unique*, solution to the decision problem. This is the best case for dominance reasoning, but things do not always work out so cleanly.

	S_1	S_2	S_3
A_1	2	3	2
A_2	2	1	0
A_3	2	3	2

Table 7.6: A decision problem in which dominance reasoning narrows the choices down without settling on a single one.

[Try it in the Decisions Lab ↗](#)

In Table 7.6, A_1 weakly dominates A_2 : $u(A_1 \wedge S_1) = u(A_2 \wedge S_1)$ (since $2 = 2$), while $u(A_1 \wedge S_2) > u(A_2 \wedge S_2)$ (since $3 > 1$) and $u(A_1 \wedge S_3) > u(A_2 \wedge S_3)$ (since $2 > 0$). We can immediately cross off A_2 , which is convenient, since it means we do not even need to check how A_2 relates to A_3 . We still need to compare A_1 and A_3 , however, and what we find is that neither dominates the other: their utilities agree at every state, so by Definition 7.11 they are equivalent. Dominance reasoning cannot, then, deliver a unique solution to this decision problem—all it can tell us is that we should not pick A_2 . That is not so bad in this case, though, since it is clear that we should pick either A_1 or A_3 , and it does not matter which.

	S_1	S_2	S_3
A_1	0	3	2
A_2	2	2	3
A_3	1	1	2

Table 7.7: A decision problem in which dominance reasoning narrows the choices down to two that are crossing.

[Try it in the Decisions Lab ↗](#)

It is harder to be so sanguine about Table 7.7. Looking at S_1 alone tells us that A_1 does not dominate either of the other choices, since it has the lowest utility at that state; and looking at S_2 alone tells us that A_1 cannot be dominated by either of the other choices, since it has the highest utility there. A_3 , however, is strongly dominated by A_2 : $u(A_2 \wedge S_1) > u(A_3 \wedge S_1)$ (since $2 > 1$), $u(A_2 \wedge S_2) > u(A_3 \wedge S_2)$ (since $2 > 1$), and $u(A_2 \wedge S_3) > u(A_3 \wedge S_3)$ (since $3 > 2$); so we can safely cross it off. That leaves us with A_1 and A_2 , and comparing them shows that $u(A_2 \wedge S_1) > u(A_1 \wedge S_1)$ (since $2 > 0$) while $u(A_1 \wedge S_2) > u(A_2 \wedge S_2)$ (since $3 > 2$): by Definition 7.12, A_1 and A_2 are crossing. All that dominance reasoning can tell us here is that we should not pick A_3 ; it is silent about how to choose between A_1

and A_2 , and unlike in Table 7.6 we cannot reassure ourselves that this silence is harmless, since A_1 and A_2 are not equivalent—it plainly matters which of the two we end up choosing.

7.3.2 Maximin

Whereas the dominance rule only tells us which choices are irrational, there are other rules for decision under ignorance which tell us which choices are rational. One of them, the maximin norm, tells us to maximize the minimum utility. To do this, we have to first figure out, for each choice, what is the minimum utility it can give us. Then we see which choice has the maximum of these minima. Of course, there may be ties. In that case, we can once again apply the maximin norm to the remaining utilities (eliminating first the minima from the previous steps). If we write $\min_i u(\varphi \wedge S_i)$ for the minimum utility that option φ yields across all states S_i —that is, φ 's worst-case utility—then the two norms are the following:

Norm 7.13. Maximin norm: Option φ is (maximin-)rational if and only if

$$\min_i u(\varphi \wedge S_i) \geq \min_i u(\psi \wedge S_i)$$

for every other option ψ .

Norm 7.14. Leximin norm: Let $u_{(1)}(\varphi) \leq u_{(2)}(\varphi) \leq \dots \leq u_{(n)}(\varphi)$ be the utilities of φ across all states, arranged from lowest to highest. Option φ is leximin-rational if and only if, for every other option ψ , either

$$u_{(k)}(\varphi) = u_{(k)}(\psi) \text{ for every } k,$$

or there is some k such that $u_{(k)}(\varphi) > u_{(k)}(\psi)$ and $u_{(j)}(\varphi) = u_{(j)}(\psi)$ for every $j < k$.

To see the maximin norm at work, consider the following decision problem, involving four options and four states. For each option we star its minimum utility, and we double-star the option whose minimum is itself the maximum among all the minima.

	S_1	S_2	S_3	S_4
A_1	5	0*	0*	2
A_2	-1*	4	3	7
A_3	6	4	4	1*
A_4^{**}	5	6	4	3*

Table 7.8: A decision problem in which the maximin norm settles on a unique option.

[Try it in the Decisions Lab ↗](#)

Going through Table 7.8 choice by choice, the minimum utility of A_1 is 0, the minimum utility of A_2 is -1 , the minimum utility of A_3 is 1, and the minimum utility of A_4 is 3. Of those four minima, 3 is the largest, and it belongs to A_4 alone: no other option has a worst case that good. So, by norm 7.13, A_4 is the uniquely maximin-rational choice.

Sometimes, though, maximin does not settle on a unique choice, because two or more choices are tied for the best worst case. When that happens, we turn to leximin, which, as we saw, tries to break the tie by looking at the second-worst utility of each of the tied choices (and, if they are tied there too, the third-worst, and so on). But even leximin can fail to settle on a unique choice, as the following decision problem, involving four choices and five states, illustrates.

	S_1	S_2	S_3	S_4	S_5
A_1	0	3	5	4	1
A_2	0	3	2	1	1
A_3	0	1	2	3	3
A_4	3	0	4	1	5

Table 7.9: A decision problem in which applying the leximin norm still leaves two choices tied.

[Try it in the Decisions Lab ↗](#)

In Table 7.9, every choice has the same minimum utility, 0, so maximin alone cannot distinguish between them. To apply leximin, we arrange each choice's utilities from lowest to highest: A_1 's become (0, 1, 3, 4, 5), A_2 's become (0, 1, 1, 2, 3), A_3 's become (0, 1, 2, 3, 3), and A_4 's become (0, 1, 3, 4, 5). Comparing these term by term, all four choices agree on the lowest utility (0) and on the second-lowest (1), so the tie is not yet broken. At the third-lowest utility, however, A_2 (with 1) and

A_3 (with 2) both fall below A_1 and A_4 (which each have 3 there), so by norm 7.14 both A_2 and A_3 are eliminated. That leaves A_1 and A_4 , and, as it turns out, their utilities sorted from lowest to highest are exactly the same: $(0, 1, 3, 4, 5)$ in both cases. So leximin narrows the field down to A_1 and A_4 , but—since their sorted utilities agree at every position—it cannot tell us which of the two to choose.

7.3.3 Minimax regret

Consider what happens when we try to apply maximin to the following table:

	S_1	S_2
A_1	\$1.50	\$1.75
A_2	\$1.00	\$10,000

Table 7.10: A decision problem in which maximin recommends the cautious option over a gamble with a much higher potential payoff.

Strictly speaking, we cannot apply any decision rule to Table 7.10 because it has amounts of money instead of utilities in the cells. But for anyone whose preferences over money is more or less normal, the corresponding utilities will have the result that the maximin norm tells us to go with A_1 because it has the maximum minimal utility possible (50 cents more than A_2). That is to say, the maximin rule is telling us to forgo an opportunity to gain \$10,000 in order to save \$0.50. Many would find that ridiculous. Those who do think that the regret you would feel if S_2 obtains and you miss out on \$10,000 is much greater than the relief you would feel if S_1 obtains and you gain \$0.5.

Let us formalize this view a bit. Let us define MAX_i to be the highest utility that any of the available options yields in state S_i —i.e., the largest number in that state’s column: $\text{MAX}_i = \max_{\varphi} u(\varphi \wedge S_i)$. Then the **regret** associated with $u(\varphi \wedge S_i)$ is the result of subtracting that value from the maximum for that state: $r(\varphi \wedge S_i) = \text{MAX}_i - u(\varphi \wedge S_i)$.

This suggests a new norm: instead of maximizing the minimum utility, minimize the maximum regret.

Norm 7.15. Minimax regret norm: Option φ is minimax-regret-rational if and only if

$$\max_i r(\varphi \wedge S_i) \leq \max_i r(\psi \wedge S_i)$$

for every other option ψ .

Returning to Table 7.10: $\text{MAX}_1 = \$1.50$ and $\text{MAX}_2 = \$10,000$. So A_1 's regret is \$0 in S_1 (since it achieves the maximum there) and \$9,998.25 in S_2 , for a maximum regret of \$9,998.25; while A_2 's regret is \$0.50 in S_1 and \$0 in S_2 , for a maximum regret of only \$0.50. By norm 7.15, then, minimax regret recommends A_2 —the reverse of what maximin recommended, and, to many people, the more sensible verdict.

To see how a regret table is built from a utility table, and how minimax regret is applied to it, consider the following decision problem, involving four choices and three states.

	S_1	S_2	S_3
A_1	4	1	6
A_2	2	5	3
A_3	6	3	1
A_4	3	4	4

Table 7.11: A decision problem used to illustrate how a regret table is built from a utility table.

[Try it in the Decisions Lab ↗](#)

We build the regret table one state at a time. In S_1 , the highest utility anyone achieves is 6 (by A_3), so $\text{MAX}_1 = 6$; in S_2 , the highest is 5 (by A_2), so $\text{MAX}_2 = 5$; and in S_3 , the highest is 6 (by A_1), so $\text{MAX}_3 = 6$. Subtracting each choice's utility from the relevant MAX_i gives us the regret table below, where we star each choice's maximum regret and double-star the choice whose maximum regret is itself the minimum among all the maxima.

	S_1	S_2	S_3
A_1	2	4*	0
A_2	4*	0	3
A_3	0	2	5*
A_4^{**}	3*	1	2

Table 7.12: The regret table corresponding to Table 7.11.

Going through Table 7.12 choice by choice, the maximum regret of A_1 is 4, the maximum regret of A_2 is 4, the maximum regret of A_3 is 5, and the maximum regret of A_4 is 3. Of those four maxima, 3 is the smallest, and it belongs to A_4 alone. So, by norm 7.15, A_4 is the uniquely minimax-regret-rational choice.

Now, maximin (and leximin) have a notable virtue that minimax regret lacks: they only need to know how the outcomes are *ranked*, not by how much one outcome is better than another. If we transform every utility in a decision problem by any function that preserves the original ranking of the outcomes—an *ordinal* transformation—maximin and leximin will always recommend the very same choice, since they only ever compare utilities using $>$, $\geq$, or $=$, and no ordinal transformation can change the truth value of any of those comparisons. Minimax regret, however, essentially involves *subtraction*, and ordinal transformations do not, in general, preserve differences between numbers. So minimax regret’s recommendation can change under a transformation that leaves maximin’s recommendation exactly where it was.

To see this, consider what happens if we square every utility in Table 7.11—a legitimate ordinal transformation, since squaring is strictly increasing on positive numbers, and hence preserves the ranking of every outcome.

	S_1	S_2	S_3
A_1	16	1	36
A_2	4	25	9
A_3	36	9	1
A_4	9	16	16

Table 7.13: The utilities of Table 7.11, each replaced by its square.

[Try it in the Decisions Lab ↗](#)

The minimum utility of A_1 is now 1, of A_2 is 4, of A_3 is 1, and of A_4 is 9—so, just as before, A_4 has the largest minimum, and maximin still uniquely recommends A_4 . Minimax regret, however, tells a different story. Now $\text{MAX}_1 = 36$, $\text{MAX}_2 = 25$, and $\text{MAX}_3 = 36$, which gives us the following regret table.

	S_1	S_2	S_3
A_1^{**}	20	24*	0
A_2	32*	0	27
A_3	0	16	35*
A_4	27*	9	20

Table 7.14: The regret table corresponding to Table 7.13.

The maximum regrets are now 24 for A_1 , 32 for A_2 , 35 for A_3 , and 27 for A_4 —so

minimax regret now uniquely recommends A_1 , the very choice it had previously ranked worst but one. Simply squaring the utilities—without changing how any outcome is ranked relative to any other—was enough to flip minimax regret’s recommendation from A_4 to A_1 , even though maximin’s recommendation did not budge. This shows that minimax regret, unlike maximin and leximin, is not a purely ordinal norm: it presupposes that differences in utility are themselves meaningful, and can be compared to one another.

Minimax regret has a second, and arguably more troubling, defect: it can violate a principle we might call *independence of irrelevant alternatives*—roughly, the idea that whether φ or ψ is recommended should not depend on which *other* choices happen to also be on the table, so long as neither of those other choices is itself the one recommended. To see minimax regret violate this principle, suppose our decision problem initially involves just two choices, A and B , and two states.

	S_1	S_2
A	9	10
B	10	8

Table 7.15: A decision problem involving two choices.

[Try it in the Decisions Lab ↗](#)

Here $\text{MAX}_1 = 10$ and $\text{MAX}_2 = 10$, so A ’s regret is $(1, 0)$, for a maximum of 1, while B ’s regret is $(0, 2)$, for a maximum of 2. By norm 7.15, minimax regret recommends A . Now suppose a third choice, C , becomes available—one that does very well in S_1 and very poorly in S_2 .

	S_1	S_2
A	9	10
B	10	8
C	15	2

Table 7.16: The same decision problem, with a third choice added.

[Try it in the Decisions Lab ↗](#)

Adding C raises MAX_1 from 10 to 15, which increases everyone’s regret in S_1 . A ’s regret is now $(6, 0)$, for a maximum of 6; B ’s regret is now $(5, 2)$, for a maximum of 5; and C ’s own regret is $(0, 8)$, for a maximum of 8. Minimax

regret now recommends B , not A —even though C , the newcomer, is not itself recommended, and nothing about A 's or B 's own prospects changed at all. It would not be strange if adding C to the menu sometimes changed the recommendation to C itself—that would just mean C turned out to be a good choice. What is strange is that merely adding C can flip the verdict between the two choices that were already on the table, even though C never had a chance of being recommended in the first place.

Finally, minimax regret can recommend choices that seem straightforwardly wrong. Consider a decision problem with a hundred states, and two choices: A_1 , which yields 100 in S_1 and nothing at all in every other state, and A_2 , which yields nothing in S_1 but 99 in every one of the other ninety-nine states.

	S_1	S_2	$\cdots$	S_{100}
A_1	100	0	$\cdots$	0
A_2	0	99	$\cdots$	99

Table 7.17: A decision problem in which minimax regret arguably gives the wrong verdict.

Intuitively, A_2 looks like the obviously better choice: it guarantees a solidly good outcome in ninety-nine out of a hundred states, while A_1 only pays off in a single state, and leaves us with nothing in all the rest. But minimax regret disagrees. $\text{MAX}_1 = 100$ and $\text{MAX}_i = 99$ for every $i > 1$. So A_1 's regret is 0 in S_1 and 99 in every other state, for a maximum regret of 99; while A_2 's regret is 100 in S_1 and 0 everywhere else, for a maximum regret of 100. Since $99 < 100$, minimax regret recommends A_1 —the choice that, intuitively, we should avoid.

There is, however, a standard reply to this last objection: given that we are dealing with a decision under ignorance, we have no idea about the relative probabilities of the 100 states. For all we know, S_1 has a 99.9999% chance of obtaining, while the remaining 99 states have almost 0 chance each. If, on the other hand, we do have at least some idea of the probabilities of the states, then we are dealing with a decision under risk, which is the topic of our next chapter.

Exercises

1. What happens to an ordinal scale when it is transformed by multiplying each number by -1? By 0? Are these acceptable transformations of the original

scale? Why or why not?

- 2. Find the maximin act for the tables below. Use leximin to break ties.**

		S_1	S_2	S_3	S_4	
a)	A_1	1	-3	5	6	Practice in the Decisions Lab ↗
	A_2	2	2	3	3	
	A_3	4	6	-10	5	

		S_1	S_2	S_3	S_4	
b)	A_1	0	1	1	3	Practice in the Decisions Lab ↗
	A_2	0	4	2	1	
	A_3	3	0	0	1	

- 3. Form the regret tables and find the minimax regret acts for the tables below.**

		S_1	S_2	S_3	
a)	A_1	7	0	4	Practice in the Decisions Lab ↗
	A_2	5	21	11	
	A_3	10	-5	-1	

		S_1	S_2	S_3	
b)	A_1	5	20	6	Practice in the Decisions Lab ↗
	A_2	-3	8	10	
	A_3	4	5	9	

8 Decisions under risk

This chapter turns to decisions under risk, where we can assign probabilities to the states, and introduces cardinal utilities, expectation, and the norm of maximizing expected utility. It then shows why that norm fails when the probability of a state depends on the choice, and replaces it with maximization of conditional expected utility.

As we saw in the previous chapter, a decision counts as a *decision under risk* when the agent cannot eliminate all but one of the states in the relevant partition, but can assign probabilities to those states. This is the case that our work on probability in Part II has prepared us for, and it is the case in which decision theory has the most to say.

8.1 Cardinal utilities

Remember the decision we discussed in relation to the minimax regret norm (Table 7.17):

	S_1	S_2	$\cdots$	S_{100}
A_1	100	0	$\cdots$	0
A_2	0	99	$\cdots$	99

Table 8.1: The decision problem of Table 7.17, repeated.

The minimax regret norm told us to go with option A_1 , because it has a regret value of 99, whereas option A_2 has a regret value of 100. However, that verdict seemed wrong: most of us are willing to let go of a single possibility of winning 100 in exchange for 99 possibilities of winning 99.

Recall that the friend of minimax regret had a reply to this worry, however. For minimax regret is advertised as a norm for decisions under ignorance. That

means that we have no idea how probable each state is. The fact that there are 99 states with an outcome of 99 for option A_2 and only 1 with an outcome of 100 for option A_1 means very little if we have no idea of how probable each state is: it may be, for all we know, that state S_1 has probability 1 of occurring, whereas the other 99 have no chance at all.

But let us now treat it as a case of decision under risk. Let us suppose, in particular, that we have equal credence for each state:

	S_1	S_2	$\dots$	S_{100}
	0.01	0.01	$\dots$	0.01
A_1	100	0	$\dots$	0
A_2	0	99	$\dots$	99

Table 8.2: The decision problem from Table 7.17, now treated as a case of decision under risk with equal credence across states.

So now that we think each state is equally likely, it is obvious that the minimax regret norm gets it wrong, right? No. If the utilities in question (the 100 for $(A_1 \wedge S_1)$ and the 99 for each of the $(A_2 \wedge S_i)$ from S_2 to S_{100}) are ordinal utilities, then we still cannot reach the conclusion that A_2 is the rational choice. For remember that utilities measured in an ordinal scale only tell us the order of the outcomes, *and nothing else*. In particular, the fact that there is only a difference of 1 between 100 and 99 means absolutely nothing for utilities measured in an ordinal scale. In fact, the very same information is transmitted by the following table:

	S_1	S_2	$\dots$	S_{100}
	0.01	0.01	$\dots$	0.01
A_1	100	0	$\dots$	0
A_2	0	1	$\dots$	1

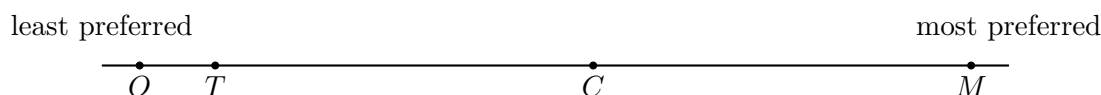
Table 8.3: The same ordinal information as before, represented with a different (order-preserving) assignment of numbers.

So, just assigning probabilities to the states is not enough to justify the intuitive thought that A_2 is the rational action instead of A_1 . What we need in order to fully justify that thought is a more fine-grained way of measuring our preferences.

What we are missing when we measure our preferences using a merely ordinal utility scale is *by how much* we prefer one outcome to another. In addition to

ordinal scales there are two other kinds of scales that are of interest to us: **interval scales** and **ratio scales**.

Interval scales can represent not just order but also the relative strength of our preferences. Thus, suppose that I prefer mate (M) to coffee (C) to tea (T) to coke (O), and that I represent my preferences with the following line:



If I measure my preferences using an ordinal scale, I need only to use numbers in ascending order from left to right. But if I want to measure my preferences using an *interval* scale, I must make sure that the relative lengths of the intervals are preserved as well. Thus, I could not assign 1 to O , 2 to T , 3 to C and 4 to M , for that would misrepresent the difference between the intervals. In general, an interval scale will be able to represent the differences in utilities, in the sense that, for any four outcomes φ, ψ, χ , and v , the preference interval between φ and ψ is greater than or equal to that between χ and v if and only if $|u(\varphi) - u(\psi)| \geq |u(\chi) - u(v)|$.

Recall that two ordinal scales are equivalent just in case they are ordinal transformations of each other. Interval scales are equivalent just in case they are **positive linear transformations** of each other:

Definition 8.1. Positive linear transformation: u' is a positive linear transformation of u if and only if there are real numbers $a > 0$ and b such that, for every outcome φ : $u'(\varphi) = a \cdot u(\varphi) + b$

Interval scales are only one type of cardinal scales. There are also ratio scales, which we use to measure weights and lengths, for instance, where there are natural zero points. For our purposes, we will need just interval scales. So, from now on, when we talk about cardinal scales you can assume that we mean interval scales, and in particular that any two cardinal scales that are positive linear transformations of each other count as equivalent.

We know that ordinal utilities come simply from your pairwise preferences among outcomes. Where do cardinal utilities come from? Before answering that question, let us see what we can do once we do have cardinal utilities to work with.

8.2 Expected value

The notion of expectation, in mathematics, is very general. Whenever we have a partition, a probability distribution over it, and a variable that takes values in each state of the partition, we have an **expectation** for that variable:

Definition 8.2. Expectation: Let $\{S_1, S_2, \dots, S_n\}$ be a partition, let P be a probability distribution over that partition, and let X be a variable that takes the value $X(S_i)$ in state S_i . Then the *expectation* of X , written $E(X)$, is:

$$E(X) = \sum_{i=1}^n P(S_i) \cdot X(S_i)$$

If you haven't seen the $\sum$ symbol before, here is what it means. $\sum$ is the capital Greek letter "sigma," and it is just shorthand for "add up." An expression like $\sum_{i=1}^n P(S_i) \cdot X(S_i)$ tells you three things: what to add up (the general term, $P(S_i) \cdot X(S_i)$), where to start ($i = 1$), and where to stop ($i = n$). To evaluate it, you plug in each value of i from 1 to n , one at a time, into the general term, and add up all the results:

$$\sum_{i=1}^n P(S_i) \cdot X(S_i) = P(S_1) \cdot X(S_1) + P(S_2) \cdot X(S_2) + \dots + P(S_n) \cdot X(S_n)$$

So the expectation of X is nothing more than a weighted sum: you take the value X takes in each state, weight it by how probable that state is, and add up the results across all the states in the partition.

Let us see some examples. Suppose, for instance, that you are thinking about how many slices of pizza you will eat for dinner. You think there is a small chance, say 10%, that you will eat just one slice, 30% chance that you will eat two slices, 50% chance that you will eat three slices, and 10% chance that you will eat four slices (you think there's no chance you will eat nothing or more than four slices). Then the expected value for the number of pizza slices that you will have for dinner is: $\frac{1}{10} \times 1 + \frac{3}{10} \times 2 + \frac{1}{2} \times 3 + \frac{1}{10} \times 4 = 2$ and $\frac{3}{5}$ slices. Or suppose that you are rolling two dice, and that you will get paid an amount of dollars equal to the sum of the faces of the die that come up. In that case, your expected gain is $\frac{1}{36} \times 2 + \frac{2}{36} \times 3 + \frac{3}{36} \times 4 + \frac{4}{36} \times 5 + \frac{5}{36} \times 6 + \frac{6}{36} \times 7 + \frac{5}{36} \times 8 + \frac{4}{36} \times 9 + \frac{3}{36} \times 10 + \frac{2}{36} \times 11 + \frac{1}{36} \times 12 = 7$ dollars.

Notice something interesting about these examples: the expected value of a variable is not, in general, the value you expect to get. Suppose that you always finish a slice of pizza once you start it. Then you do not expect to eat 2 and $\frac{3}{5}$

slices, but that is the expected number of pizza slices you will eat in the technical sense of "expected." Similarly, you do not really expect to get paid 7 dollars in the dice game (there is only a $\frac{6}{36}$ probability of that happening), and yet that is the expected monetary value of playing that game. If the expected value is not the value you expect to obtain, then what is it?

One answer to that question is that the expected value is the average value you should expect to obtain if you do the same thing over and over. For instance, if you play the die game a number of times, on average you will get 7 dollars per game, and so the higher the number n of games you play, the closer your average gain per game will be to 7.¹ Expected value is, after all, a weighted average. But suppose you will play the game just once, and so this line of thought doesn't apply. Does expected value have no meaning though? Arguably, it still does. You can think of expected value as a way of balancing risks: sure, you may be unlucky to win only 2 dollars, but you may also get lucky and win 12. Most likely, though, you'll win something in between. Here a weighted average does not seem such a bad idea.

7 dollars is the *expected monetary value* of our dice game. Likewise, the expected monetary value of a ticket in the "SuperEnalotto" we talked about in Chapter 6 is, if we suppose the jackpot stands at 100,000,000 euros and we set aside the smaller secondary prizes: $\frac{1}{622,614,630} \times 100,000,000 + \frac{622,614,629}{622,614,630} \times 0 \approx 0.16$ euros—a small fraction of the 1 euro it costs to buy the ticket.²

Does that mean that anyone who buys a ticket in the SuperEnalotto is simply irrational? Not necessarily. Suppose, for instance, that one needs millions of euros in order to pay ransom for the kidnapping of one's pet, and that one simply does not have that much money. If you care a lot about your pet, you might think it's worth paying 1 euro for the minuscule possibility of saving your pet. Or perhaps some people simply enjoy the thrill of playing the lottery. In both of these cases, the utility associated with buying a ticket cannot simply be equated with its expected monetary value.

8.3 Where do cardinal utilities come from?

Recall that ordinal utilities come simply from your preference order among outcomes. Where do cardinal utilities come from? Not from the expected monetary value of the outcomes, as we just saw. But then, from where?

¹This is the Law of Large Numbers.

²Taking into account all the minor prizes and the jackpot rollover complicates things a bit, but still the expected monetary value of a ticket is around 0.6 euros.

Suppose that you like coffee more than tea, and tea more than coke. So far, that just gives us a ranking. But if you can answer some additional questions, we can transform that ranking into an interval scale. The kinds of questions you have to answer have the following form: what do you prefer, tea outright, or a lottery ticket with 50% chance of winning a coffee and 50% chance of winning a coke? Suppose you answer that you prefer the lottery. In that case, the difference between your utility for coffee and your utility for tea is larger than the difference between your utility for tea and your utility for coke: your preference for coffee over tea is stronger than your preference for tea over coke. If, on the other hand, you prefer tea outright to the lottery, then that means that the difference between your utility for tea over coke is larger than the difference between your utility for coffee over tea: your preference for tea over coke is stronger than your preference for coffee over tea. In each of these cases, we still don't know how strong the difference in preferences is. For that, we need to ask you about lotteries with other chances for coffee and coke, until we hit one lottery such that you are indifferent between that lottery and tea outright. Finally, if you answer that you are indifferent between tea and the 50-50 lottery, then that means that the difference between your utility for coffee and your utility for tea is exactly the same as the difference between your utility for tea and your utility for coke: tea sits right in the middle between coffee and coke for you.

What we did for tea we can do for every other outcome. To fix the scale we also need two anchor points, which we are free to stipulate: say, utility 0 for coke and utility 1 for coffee, so that indifference between tea and the 50-50 lottery gives tea a utility of 0.5. In that way, we can construct a cardinal utility scale for you. This method assumes that you value a lottery by weighting the utilities of its prizes by their probabilities; otherwise we could not tell whether you prefer tea outright because the difference between tea and coke is large, or merely because you dislike risk. Von Neumann and Morgenstern showed that this assumption holds when your preferences over lotteries meet certain plausible conditions, which we will not study here.

8.4 Maximization of expected utility

Now that we have utilities measured in a cardinal scale, we can make sense of the *expected utility* of a choice φ :

Definition 8.3. Expected utility: Let $\{S_1, S_2, \dots, S_n\}$ be a partition of states, let P be a probability distribution over that partition, and let u be a cardinal utility function. Then the *expected utility* of a choice φ , written $EU(\varphi)$, is:

$$EU(\varphi) = \sum_{i=1}^n P(S_i) \cdot u(\varphi \wedge S_i)$$

And, with all that preamble out of the way, we are now in a position to define perhaps the most famous decision norm of them all, maximization of expected utility:

Norm 8.4. Maximization of expected utility (MEU): when making a decision under risk with cardinal utilities, perform the action which maximizes expected utility. A choice maximizes expected utility if and only if:

$$EU(\varphi) \geq EU(\psi)$$

for every other option ψ .

Notice that the maximization of expected utility norm allows for ties.

Now that we have to represent probabilities of states in addition to utilities for outcomes, we need to modify our decision matrices to include this extra information. Let us go back to our park/movies example from last chapter, and this time let us assume that we have a 0.7 credence that it will rain and the corresponding 0.3 credence that it will not rain. The matrix for this example is the following:

	R $\frac{7}{10}$	$\neg R$ $\frac{3}{10}$
P	1	4
M	3	2

Table 8.4: The decision problem of whether to go to the park (P) or to the movies (M), now treated as a case of decision under risk, with credence $\frac{7}{10}$ that it will rain and $\frac{3}{10}$ that it will not.

Let us now find out the expected utility for each option:

$$EU(P) = 1 \times \frac{7}{10} + 4 \times \frac{3}{10} = 1.9$$

$$EU(M) = 3 \times \frac{7}{10} + 2 \times \frac{3}{10} = 2.7$$

The result is that going to the movies is the choice that maximizes expected utility, and thus the rational choice according to the MEU norm.

We should notice an interesting fact about the interaction between the dominance norm and the MEU norm. Even though we talked about dominance in connection with decisions under ignorance, dominated choices will never maximize expected utility. This is because if a choice is dominated then its expected utility is guaranteed to be less than that of the choice that dominates it. Moreover, if there is just one non-dominated choice, then it will maximize expected utility. In that sense, the dominance norm and the MEU norm cannot enter into conflict. And this means that if in a decision under risk just one choice is non-dominated, then we do not need to bother with calculating expected utility. Consider, for instance, what happens when we modify the example so that going to the park dominates going to the movies:

	R	$\neg R$
	$\frac{7}{10}$	$\frac{3}{10}$
P	5	4
M	3	2

Table 8.5: A modified version of the park/movies decision problem in which going to the park (P) dominates going to the movies (M).

It is easy to see why if there is a unique dominant choice then it will maximize expected utility: because in the calculation of expected utility, the dominant choice has larger utilities no matter what state occurs, and the utilities are weighted by the same probability for each option:³

$$EU(P) = 5 \times \frac{7}{10} + 4 \times \frac{3}{10} = 4.7$$

$$EU(M) = 3 \times \frac{7}{10} + 2 \times \frac{3}{10} = 2.7$$

³If there is a unique *weakly* dominant choice, then it will have larger utilities for at least some state, and no lower utilities in any state, with the same result.

8.5 When states depend on choices: maximization of conditional expected utility

The fact that dominance and the MEU norm never conflict gives rise to a problem. A lazy student has two options: study for the class, or not study. There are two states the world could be in: either he passes the class or he doesn't. He prefers passing to not passing, but, for every state, he prefers not studying to studying. His utilities can be whatever you want compatible with this. Let us assume they are as follows:

Dictionary:

S = The student studies.

P = The student passes the class.

Utilities:

$$u(\neg S \wedge P) = 7$$

$$u(\neg S \wedge \neg P) = 2$$

$$u(S \wedge P) = 6$$

$$u(S \wedge \neg P) = 1$$

To complete the decision matrix we would have to know what credence the student has that he will pass the class. But, on second thought, it doesn't really matter, for not studying dominates studying for this student. Whatever probability n he assigns to passing, his matrix will look like this:

	P n	$\neg P$ $1 - n$
S	6	1
$\neg S$	7	2

Table 8.6: The lazy student's decision matrix.

Notice that the expected utilities are as follows:

$$EU(S) = 6n + 1(1 - n) = 5n + 1$$

$$EU(\neg S) = 7n + 2(1 - n) = 5n + 2$$

Thus, as we know, dominance and the MEU norm give the same result: the student should not study.

But this is obviously the wrong result, for the model of the decision is missing one crucial fact: studying will increase the probability that the student will pass, whereas not studying will decrease it. So far, we have been dealing with the unconditional probabilities of the states, which means that we assign any outcome under a given state the same probability. In this case, passing the class has the same probability no matter whether the student studies or not. Dealing with these unconditional probabilities is fine when the probability of the states does not vary with the choices the agent makes. For instance, going to the movies or going to the park does not affect the probability that it will rain, and so for that kind of decision problem it is ok to assign unconditional probabilities to the states. But for the kind of problem where the choices do affect the probability of the states, as is the case with the lazy student, we need to consider, not the unconditional probability that the state will occur, but rather the conditional probability that the state will occur given that a certain action is chosen.

When dealing with unconditional probabilities it is fine to have them just under the label for the state, because the same probability can be used for the whole column, but when dealing with conditional probabilities we need to consider potentially different conditional probabilities for different outcomes under the same column, and so we will write the probabilities in the cells. The utilities will go in the upper-left corner of the cell and the probabilities in the lower right. Thus, let us assume that the conditional probability that the student will pass given that he studies is 0.7, whereas the conditional probability that the student will pass given that he doesn't study is 0.1.⁴ Under these assumptions, the matrix for this decision problem looks like this:

	P	$\neg P$
S	6 0.7	1 0.3
$\neg S$	7 0.1	2 0.9

Table 8.7: The lazy student's decision matrix using conditional probabilities. In each cell, the utility appears in the upper-left corner and the conditional probability of the state given the choice appears in the lower-right corner.

⁴Notice that whereas conditional probabilities don't have to add up to 1 under the columns, they do have to add up to 1 across the rows, and thus the conditional probability that the student will not pass given that he studies is 0.3, and the conditional probability that the student will not pass given that he does not study is 0.9.

Now, how do we calculate the expected utility of each choice? Well, instead of using the unconditional probability of the state to multiply the utility of each outcome, we use the conditional probability of the state given the choice:

Definition 8.5. Conditional expected utility: Let $\{S_1, S_2, \dots, S_n\}$ be a partition of states, let P be a probability function, and let u be a cardinal utility function. Then the *conditional expected utility* of a choice φ relative to P , written $CEU(\varphi)$, is:

$$CEU(\varphi) = \sum_{i=1}^n P(S_i | \varphi) \cdot u(\varphi \wedge S_i)$$

Corresponding to this definition of conditional expected utility, we have the following maximization of conditional expected utility norm:

Norm 8.6. Maximization of conditional expected utility: When making a decision under risk with cardinal utilities where the probability of the states depends on the choice, perform the action which maximizes conditional expected utility. A choice maximizes conditional expected utility if and only if:

$$CEU(\varphi) \geq CEU(\psi)$$

for every other option ψ .

Applied to the lazy student case, the MCEU norm gives the intuitively right result that the student should study:

$$CEU(S) = 0.7 \times 6 + 0.3 \times 1 = 4.2 + 0.3 = 4.5$$

$$CEU(\neg S) = 0.1 \times 7 + 0.9 \times 2 = 0.7 + 1.8 = 2.5$$

Since $CEU(S) = 4.5 > 2.5 = CEU(\neg S)$, by MCEU, the student should study.

So now that we have two different norms for decisions under risk, MEU and MCEU, which one do we apply when? The straightforward answer is that we should apply MEU when the states are probabilistically independent of the choices, and MCEU otherwise. As we will see in the next chapter, however, this simple answer gives rise to some complicated issues.

Exercises

For each of the decision problems below, write down the corresponding decision matrix and then apply MEU or MCEU as appropriate.

1. A basketball player must decide whether to practice free throws every day before a big game (T) or skip practice ($\neg T$). The relevant states are whether he makes the game-winning free throw (W) or misses it ($\neg W$). Practicing raises the conditional probability of making the shot: $P(W | T) = 0.8$ and $P(W | \neg T) = 0.5$. His utilities are $u(T \wedge W) = 10$, $u(T \wedge \neg W) = 3$, $u(\neg T \wedge W) = 10$, and $u(\neg T \wedge \neg W) = 4$.
2. An investor is deciding whether to put her savings into stock A or stock B . The market will either go up (M) or down ($\neg M$) next year. She assigns a credence of 0.55 to the market going up. Her utilities are $u(A \wedge M) = 8$, $u(A \wedge \neg M) = 2$, $u(B \wedge M) = 5$, and $u(B \wedge \neg M) = 4$.
3. A gambler must choose between betting on red (A_R) or betting on black (A_B) on a single spin of a roulette wheel. The relevant states are that the wheel lands red (R) or lands black (B), and she believes the wheel has a 0.52 chance of landing red and a 0.48 chance of landing black. If she bets red and it lands red, she wins \$10; if she bets red and it lands black, she loses \$5. If she bets black and it lands black, she wins \$8; if she bets black and it lands red, she loses \$4. (Assume her utilities equal the dollar amount won.)
4. A woman is deciding whether to exercise regularly (E) or remain sedentary ($\neg E$) this year. The relevant states are whether she gets seriously ill (I) or stays healthy ($\neg I$). Exercising lowers the conditional probability of illness: $P(I | E) = 0.1$ and $P(I | \neg E) = 0.3$. Her utilities are $u(E \wedge I) = 2$, $u(E \wedge \neg I) = 8$, $u(\neg E \wedge I) = 1$, and $u(\neg E \wedge \neg I) = 9$.
5. You are deciding whether to bring an umbrella (U) or leave it at home ($\neg U$) on your way to work. There are two possible states: it rains (R) or it doesn't ($\neg R$). You assign a credence of 0.4 to rain. Your utilities are $u(U \wedge R) = 5$, $u(U \wedge \neg R) = 3$, $u(\neg U \wedge R) = 1$, and $u(\neg U \wedge \neg R) = 6$.
6. A driver must decide whether to put her phone away while driving (C) or keep glancing at it ($\neg C$). The relevant states are whether she gets into a serious accident (A) or not ($\neg A$). Putting the phone away lowers the conditional probability of an accident: $P(A | C) = 0.02$ and $P(A | \neg C) = 0.15$. Her utilities are $u(C \wedge A) = 1$, $u(C \wedge \neg A) = 9$, $u(\neg C \wedge A) = 0$, and $u(\neg C \wedge \neg A) = 10$.
7. A person is deciding whether to get a flu vaccine (V) or skip it ($\neg V$) this winter. The relevant states are whether she catches the flu (F) or not ($\neg F$).

Getting vaccinated lowers the conditional probability of catching the flu: $P(F | V) = 0.05$ and $P(F | \neg V) = 0.25$. Her utilities are $u(V \wedge F) = 4$, $u(V \wedge \neg F) = 9$, $u(\neg V \wedge F) = 2$, and $u(\neg V \wedge \neg F) = 10$.

8. A salesperson is deciding whether to spend an extra hour each day cold-calling potential clients (H) or use that hour for something else ($\neg H$). The relevant states are whether she closes a major deal this month (D) or not ($\neg D$). Cold-calling raises the conditional probability of closing a deal: $P(D | H) = 0.4$ and $P(D | \neg H) = 0.1$. Her utilities are $u(H \wedge D) = 10$, $u(H \wedge \neg D) = 4$, $u(\neg H \wedge D) = 12$, and $u(\neg H \wedge \neg D) = 5$.

9 Newcomb's problem

This chapter examines Newcomb's problem, a decision under risk in which the states are probabilistically dependent on the choices but not causally dependent on them, so that it is controversial whether MEU or MCEU applies. It presents the arguments for one-boxing and for two-boxing, including the "why ain'tcha rich" argument, the two-boxer's lack-of-opportunities reply, and the transparent variant of the problem.

9.1 Introduction

In the last chapter we learned that there are two different kinds of decisions under risk: those where the states are probabilistically independent of the choices and those where they are not. The difference is important because different norms apply depending on the type of decision we are faced with. If the states are independent of the choices, then the MEU norm applies, whereas if they are not, the MCEU applies.

In this chapter we are going to examine a kind of decision where the states are probabilistically dependent on the choices, but nevertheless there is a controversy as to whether we should apply MEU or MCEU. This is because the probabilistic dependence is not accompanied (and explained) by a causal dependence. In cases like that, it is not obvious which norm we should apply.

9.2 The case

Predictor is an agent who is *extremely good* at predicting people's choices. He is not infallible. In fact, Predictor's track record shows that he has made the right prediction about people's choices around 90% of the time. It is crucial to realize that, although Predictor is very good at predicting choices, he is not a supernatural entity that can travel in time towards the future to see what your choice is. He is

simply a very good student of human psychology, and has been predicting choices (and learning from whether his predictions are true or not) for a long time. This time, Predictor will predict your choice in Newcomb's problem.¹

In front of you there are two boxes: a transparent one and an opaque one. You can see that there are \$1,000 in the transparent box, but you cannot see the contents of the opaque box. You can either pick both boxes and keep their contents, or pick just the opaque box.

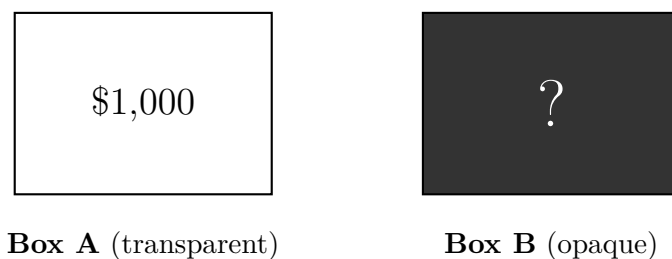


Figure 9.1: The setup in Newcomb's problem. Box A is transparent: you can see that it contains \$1,000. Box B is opaque: you cannot see whether it is empty or contains a large sum of money. You may take both boxes, or Box B alone.

If the foregoing was all you knew about your choices, then you should obviously pick both boxes (we will henceforth call picking both boxes "two-boxing") instead of just the opaque one ("one-boxing"). For, if you pick both boxes, then you get to keep whatever is inside the opaque box (if anything) *plus* the extra \$1,000 in the transparent box. This is *not* Newcomb's problem, and there are no controversies as to what you should do in this case.

But the twist is that you know something else about the case. You know that the contents of the opaque box were determined *yesterday* by Predictor, according to the following method: he first made a prediction regarding your choice today; then, if he predicted that you were going to two-box, he left the opaque box empty; but if he predicted that you were going to one-box, he put a million dollars inside the opaque box. The decision is represented in the following table:

¹Newcomb was a physicist friend of the philosopher Robert Nozick. Nozick popularized the problem.

	Predicted one-boxing	Predicted two-boxing
One-box	\$1,000,000	\$0
Two-box	\$1,001,000	\$1,000

Table 9.1: The payoffs in Newcomb's problem, according to your choice (rows) and Predictor's prediction (columns).

Table 9.1 is not actually a decision matrix, because we have neither the subject's utilities nor his or her credences. But we can assume that the subject's utilities are more or less linear with the amount of money to be gained in each outcome, so that the decision matrix with utilities looks something like the following:²

	Predicted one-boxing	Predicted two-boxing
One-box	1,000,000	0
Two-box	1,001,000	1,000

Table 9.2: An approximate decision matrix for Newcomb's problem, treating the subject's utility for each outcome as equal to the number of dollars gained.

Now that we have utilities, we can see that two-boxing dominates one-boxing. If this were a decision under ignorance, then, two-boxing would be the only rational choice. Moreover, if this were a decision where the states are probabilistically independent of the choices, so that MEU applies, then two-boxing will also be the only rational choice. Let n be your credence that Predictor predicted that you were going to one-box (so that $1 - n$ is your credence that he predicted you were going to two-box). Then:

$$EU(\text{One-box}) = 1,000,000n + 0 \times (1 - n) = 1,000,000n$$

$$EU(\text{Two-box}) = 1,001,000n + 1,000 \times (1 - n) = 1,000,000n + 1,000$$

Since $EU(\text{Two-box}) = EU(\text{One-box}) + 1,000$, two-boxing has the greater expected utility no matter what value n takes—which is just what we should expect, given that two-boxing dominates one-boxing.

But Newcomb's case is neither a case of decision under ignorance nor one where the states are independent of the choices. We said that Predictor is 90% accurate. We will take this to mean that he is equally reliable whichever choice you make: if

²Given the diminishing marginal utility of money, this assumption is wrong, but harmless in this context.

you one-box, there is a 90% probability that he predicted it, and if you two-box, there is a 90% probability that he predicted that. That means that the conditional probabilities of the states given the choices are not the same:³

Dictionary:

O = You one-box.

D = Predictor predicts that you one-box.

$$\begin{aligned} P(D | O) &= 0.9 & P(\neg D | O) &= 0.1 \\ P(D | \neg O) &= 0.1 & P(\neg D | \neg O) &= 0.9 \end{aligned}$$

If we add this information to our decision matrix, the result is the following:

	D	$\neg D$
O	1,000,000 0.9	0 0.1
$\neg O$	1,001,000 0.1	1,000 0.9

Table 9.3: The decision matrix for Newcomb's problem, with the conditional probability of each state given each choice in the lower right of each cell.

So Newcomb's case is a problem of decision under risk where the states are probabilistically dependent on the choices. Given the reliability of Predictor, the probability that he predicted that you are going to one-box given that in fact you will one-box is higher than the unconditional probability that he predicted that you are going to one-box (and the same goes for two-boxing). It seems, then, that the norm which applies here is MCEU rather than MEU:

$$CEU(O) = 0.9 \times 1,000,000 + 0.1 \times 0 = 900,000 + 0 = 900,000$$

$$CEU(\neg O) = 0.1 \times 1,001,000 + 0.9 \times 1,000 = 100,100 + 900 = 101,000$$

Since $CEU(O) = 900,000 > 101,000 = CEU(\neg O)$, by MCEU, you should one-box.

³We are assuming that you two-box if and only if you do not one-box.

9.3 The initial arguments

There are very good arguments for one-boxing. The problem is that there are very good arguments for two-boxing as well.

The initial argument for one-boxing consists simply in pointing out the analogy between Newcomb's case and the lazy student case. Just as in the lazy student case we should not care about the fact that not studying dominates studying because studying raises the probability of passing, so too in Newcomb's case we should not care that two-boxing dominates one-boxing because one-boxing raises the probability that Predictor predicted that we were going to one-box, which in turn raises the probability of us ending up with \$1,000,000.

The two-boxer retorts that there is a crucial difference between the lazy student's case and Newcomb's case. In the lazy student's case, studying raises the probability of passing because studying has a causal effect on how well you do in the class. On the other hand, one-boxing raises the probability that Predictor predicted that you were going to one-box, but *not* because it has a causal effect on what Predictor predicts. We established that Predictor does not travel in time, and that what you do in the future cannot causally affect what he predicts today. Rather, what is going on is that if you have a propensity to one-box, then Predictor will have picked that up when he studied you, and so if you in fact one-box it is highly likely that Predictor predicted that you were going to one-box. But the prediction is already made by the time you have to choose, and so your choice cannot have a causal influence on how much money you get. Either the million dollars is already in the opaque box or it isn't, and nothing you do now can change that. Therefore, you should two-box.

This two-boxing reply to the initial one-boxing argument can be buttressed by the different causal structures that the cases have. In the lazy student's case, studying causes the student to do better in the class, and that is why it raises the probability of passing. In Newcomb's case, one-boxing raises the probability that Predictor predicted one-boxing because the prediction and the agent's choice have a common cause—the agent's propensity to one-box:

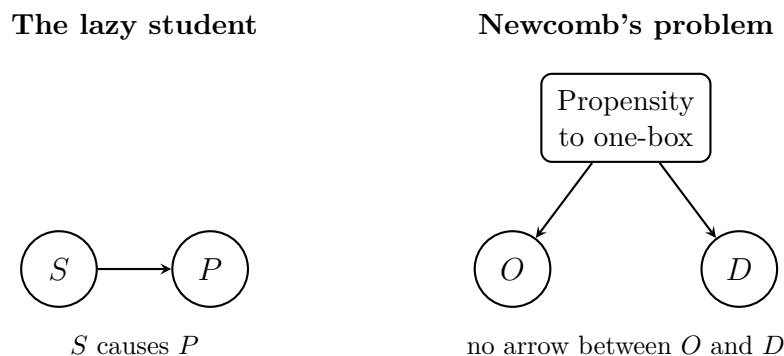


Figure 9.2: The causal structure of the lazy student's decision (left) compared to that of Newcomb's problem (right). Studying (S) directly causes passing (P). By contrast, one-boxing (O) and Predictor's prediction that you will one-box (D) are correlated only because they share a common cause—your propensity to one-box—and neither one causes the other.

It may help to make the two-boxer point more vivid to slightly change the way the decision goes, without actually changing the payoff structure of the problem. Suppose that, as you enter the room, you are given \$1,000 and the opaque box. Your decision is simply whether you want to give the \$1,000 back or not. The contents of the opaque box were determined, as always, by Predictor: he put \$1,000,000 in it if and only if he predicted you were going to give back the \$1,000. Structuring the problem this way dramatizes the two-boxer's point: giving back the \$1,000 has no effect whatsoever on the contents of the opaque box, which was fixed yesterday.⁴

To recap: the one-boxer's initial argument is that MCEU is the right norm to apply to the Newcomb's case, and that it recommends one-boxing. The two-boxer's reply to this is that MCEU is wrong when applied this way. According to the two-boxer, we should apply MCEU only when the probabilistic dependence is underwritten by a causal dependence, because we should choose the option that has the best chance of *causing* a favorable outcome. Given that Newcomb's problem does not have the right causal structure, one-boxing will not have any causal effect on what is inside the opaque box: its only relevant causal effect will be that we have \$1,000 less than what we could have had.

⁴In addition, presenting the problem in this way triggers the *loss aversion* reactions: the tendency for the pain of losing something to feel more intense than the pleasure of getting the same thing.

9.4 “Why ain’cha rich,” lack of opportunities, and transparent Newcomb

One way in which one-boxers can push back against two-boxers is as follows. Let us assume that everyone involved thinks that more money is better than less.⁵ If we compare the population of one-boxers to that of two-boxers, there is a striking difference: around 90% of one-boxers are millionaires, whereas only around 10% of two-boxers are. So one-boxers taunt two-boxers: if you are so smart, why ain’cha rich?

Two-boxers must concede this point: one-boxers really do have more money than two-boxers. But two-boxers point out that neither the one-boxer’s riches nor the two-boxer’s poverty are of their own doing. Pretty much every one-boxer could have been richer than they are, to the tune of an extra \$1,000. Pretty much no two-boxer could have had more than \$1,000. This is because most one-boxers were choosing between \$1,000,000 and \$1,001,000, and foolishly left the \$1,000 on the table, whereas most two-boxers were choosing between \$1,000 and nothing and rationally chose the \$1,000.⁶ We can call this the “lack-of-opportunities” reply to the one-boxer’s *why ain’cha rich* argument.

We can dramatize the lack-of-opportunities reply by imagining yet another small variant of the case that doesn’t affect the payoff structure. Suppose that there are two rooms. In room 1 the opaque box contains \$1,000,000, whereas in room 2 it is empty. Predictor then makes his predictions as follows: if it predicts that the agents will one-box, then he sends them to room 1, whereas if he predicts that they will two-box, he sends them to room 2. The two-boxer can then say that he is not rich because he was sent to room 2, where there were no riches to be had, whereas the one-boxer is rich (despite himself) because he was sent to room 1, and it is impossible to walk out of room 1 without \$1,000,000.

⁵Strictly speaking this assumption is not necessary, for we can replace money with whatever it is that the agent values.

⁶In addition, a few one-boxers got nothing, and a few two-boxers got \$1,001,000.

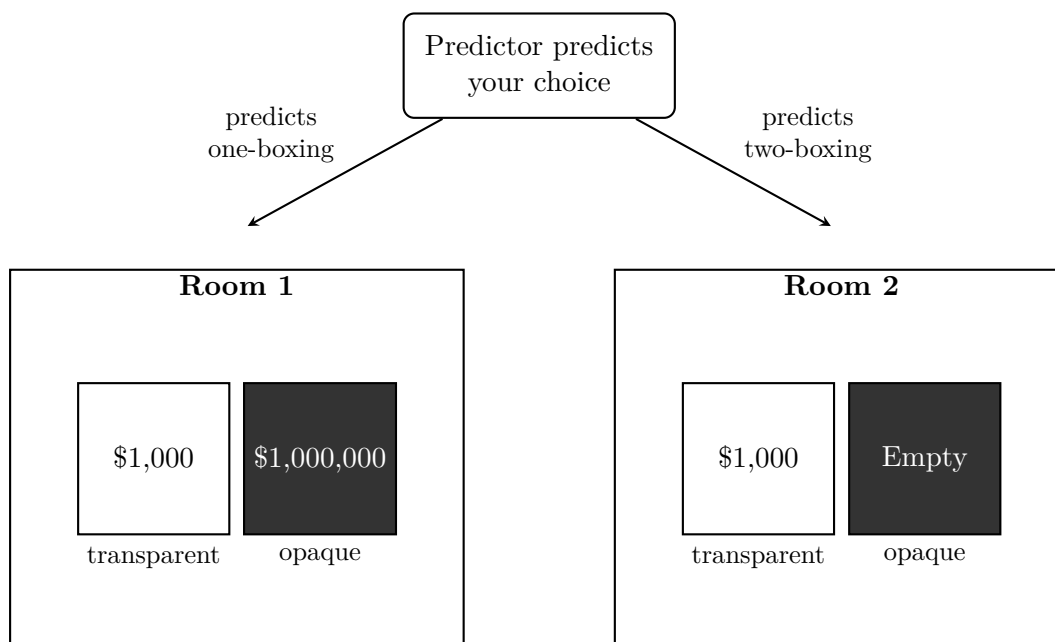


Figure 9.3: The two-room variant of Newcomb's problem. Predictor sends you to Room 1 if he predicts you will one-box, and to Room 2 if he predicts you will two-box. Either way, once you are in a room the payoff structure is exactly as before: a transparent box with \$1,000, and an opaque box whose contents were fixed when you were assigned to the room. It is impossible to leave Room 1 with less than \$1,000,000, and impossible to leave Room 2 with more than \$1,000.

The one-boxer must concede that which room you are sent to determines what opportunities you have. But it's not as if it is completely out of your hands which room you are sent to. You can make a determination to one-box, in which case you will most likely be sent to room 1! To the two-boxers' complaint that they don't have the opportunity to be rich, one-boxers can reply: "make your own opportunities!"

And now it is the turn of the two-boxer to concede something: he too would be a one-boxer if only he could. From the perspective of the two-boxer, Predictor is rewarding the irrational choice of one-boxing. Given that this choice is rewarded with \$1,000,000, it makes sense to try to be a one-boxer. The one-boxer of course agrees that it makes sense to try to be a one-boxer, but he has two questions for the two-boxer: first, why doesn't this mean that one-boxing is, after all, the rational choice; second, what is it with all this talk about *trying* to be a one-boxer, where it is the easiest thing in the world: you just have to one-box?

The two-boxer's reply to the first question starts by making a general point: it

can sometimes be rational to try to do something irrational. Suppose, for instance, that someone offers you the following deal: he will give you \$1,000 today if you decide now to do something irrational tomorrow. Notice well: the \$1,000 does not require you to do something irrational tomorrow; you are only required to decide *today* to do something irrational tomorrow. Now, there is a question as to whether you can genuinely decide to do something irrational tomorrow. For, when tomorrow comes, either you will have won the \$1,000 or not. If you did, no reason to do something irrational. And if you didn't, then it would only add insult to injury to do something irrational.⁷ But in any case, it would be good for you if you could decide to do something irrational. This is analogous to what the two-boxer thinks happens with one-boxing. It is an irrational thing to do, but given that it is rewarded it makes sense to try to be a one-boxer. When the time comes, however, you know that, no matter which room you are in, you are better off taking the \$1,000. Hence the difficulty in carrying through the intention to one-box: faced with the actual decision, the two-boxer finds it impossible not to take the free \$1,000—not out of respect for some abstract notion of rationality, but because they are \$1,000 for free.

Now, if only there were a one-boxing pill that one can take in full view of Predictor, that would change things. This one-boxing pill forces one to be a one-boxer. The two-boxer would happily take such a pill before Predictor makes his prediction, for then he would be sent to room 1. Now, once in room 1, the pill-taker will do what the two-boxer considers to be irrational: he will one-box. But being irrational in room 1 beats being rational in room 2—although it doesn't make one-boxing rational in either room.

The one-boxer can point out that the two-boxer is then committed to denying the following principle: if it is rational to force oneself to φ , then it is rational to φ . For the two-boxer has conceded that it is rational to force oneself (if possible) to be a one-boxer, and yet insists that it is irrational to be a one-boxer. But isn't it absurd to deny that principle?

The two-boxer, of course, doesn't think it is absurd, as just explained. But, more to the point, there is an argument that even the one-boxer must deny the principle. We must now consider a genuine variation on Newcomb's problem: *transparent Newcomb*. Transparent Newcomb is a genuine variation in the sense that, unlike the two rooms variations, is not simply the same decision problem as the Newcomb case differently described, but rather has a different payoff structure. Transparent

⁷This is a version of Kavka's "toxin puzzle".

Newcomb is just like the Newcomb case except that the two boxes are transparent: you can see whether or not the \$1,000,000 is in the second box. So you will face one of two decision problems, depending on which room you are sent to: a choice between \$1,000 and nothing, or a choice between \$1,001,000 and \$1,000,000.

Room 1		Room 2	
\$1,000	\$1,000,000	\$1,000	\$0
Box A	Box B	Box A	Box B
Two-box: \$1,001,000		Two-box: \$1,000	
One-box: \$1,000,000		One-box: \$0	

Figure 9.4: Transparent Newcomb. Both boxes are now transparent, so you can see whether Box B contains the \$1,000,000 before choosing. Depending on which room you were sent to, you face one of two decision problems: a choice between \$1,001,000 and \$1,000,000 (Room 1), or a choice between \$1,000 and nothing (Room 2). In either case, two-boxing is visibly the dominant choice.

In either case, MCEU gives the same verdict as MEU: two-boxing is the only rational option.

	D	$\neg D$
O	1,000,000 1	0 0
$\neg O$	1,001,000 1	1,000 0

Table 9.4: The decision matrix in Room 1 (Box B contains \$1,000,000).

$$CEU(O) = 1 \times 1,000,000 + 0 \times 0 = 1,000,000$$

$$CEU(\neg O) = 1 \times 1,001,000 + 0 \times 1,000 = 1,001,000$$

Since $CEU(\neg O) = 1,001,000 > 1,000,000 = CEU(O)$, by MCEU, you should two-box.

	D	$\neg D$
O	1,000,000 0	0 1
$\neg O$	1,001,000 0	1,000 1

Table 9.5: The decision matrix in Room 2 (Box B is empty).

$$CEU(O) = 0 \times 1,000,000 + 1 \times 0 = 0$$

$$CEU(\neg O) = 0 \times 1,001,000 + 1 \times 1,000 = 1,000$$

Since $CEU(\neg O) = 1,000 > 0 = CEU(O)$, by MCEU, you should two-box.

The difference between the original Newcomb case and the two possible transparent Newcomb cases you may face is that the latter have extreme and independent probabilities: because both boxes are transparent, you can see what is in them, and therefore your credences for what is in them are the same whether you one-box or not. The decision in either of the transparent Newcomb cases is whether or not you want a free \$1,000—or, more precisely, whether or not you choose to get what you want or not. Now, even though almost everybody, one-boxers included, agrees that two-boxing is the only rational option in transparent Newcomb,⁸ there are patently irrational *transparent one-boxers*, who one-box in transparent Newcomb. They are sent to room 1, of course, and there they leave \$1,000 on the table. However, transparent one-boxers are to one-boxers what one-boxers are to two-boxers: they win millions from transparent Newcomb, whereas one-boxers gain merely thousands. If there were a transparent one-boxing pill, one-boxers as well as two-boxers would happily take it, for being irrational in room 1 beats being rational in room 2, for transparent Newcomb as much as for the original Newcomb case. Which means that even one-boxers are committed to denying the principle that if it is rational to force oneself to φ , then it is rational to φ : they think it is rational to force oneself to be a transparent one-boxer, even though they think it is irrational to be a transparent one-boxer. Therefore, one-boxers cannot complain that two-boxers are committed to denying that principle.

⁸Not quite everybody: some decision theorists, such as proponents of functional decision theory, hold that one should one-box even in transparent Newcomb, as in the case known as Parfit's hitchhiker. We set that view aside here.

Exercises

For each of the following, construct the relevant decision matrix (or matrices) using conditional probabilities, and compute CEU for one-boxing and two-boxing to determine what MCEU recommends.

1. Redo the original Newcomb's problem (\$1,000 in the transparent box, \$1,000,000 in the opaque box) assuming Predictor is only 70% reliable instead of 90%. Does MCEU still recommend one-boxing?
2. Redo the original Newcomb's problem keeping Predictor's reliability at 90%, but suppose the opaque box contains \$10,000,000 instead of \$1,000,000 (the transparent box still contains \$1,000). Does MCEU's recommendation change?
3. Predictor's reliability can be represented by a number r : the probability that his prediction matches your actual choice, so that $P(D|O) = P(\neg D|\neg O) = r$ and $P(D|\neg O) = P(\neg D|O) = 1 - r$. Keeping the \$1,000 and \$1,000,000 payoffs fixed, find the value of r at which $CEU(O) = CEU(\neg O)$. For values of r above this tipping point, what does MCEU recommend? What about below it?
4. Suppose instead that Predictor's reliability is fixed at 90%, the opaque box still contains \$1,000,000 or nothing, but the amount of money in the transparent box is increased from \$1,000 to some larger amount x . Find the value of x at which $CEU(O) = CEU(\neg O)$.

10 Games

This chapter turns to games, decisions in which the outcome depends on the choices of other agents, and shows how to represent them as matrices and trees under the standard assumptions of game theory. It then introduces the main tools for solving them (iterated deletion of dominated strategies, Nash equilibrium and saddle points, and backwards induction) and examines the Prisoner's Dilemma, its connection to Newcomb's problem, and the centipede game.

10.1 Games as interactive decisions

So far, every decision we have considered has been made against a partition of *states of nature*: possibilities that, for the most part, obtain or fail to obtain independently of what the agent decides to do. Whether it rains has nothing to do with whether you go to the park or to the movies; whether a given company's stock rises has nothing to do with which stocks you bought. This independence is what let us treat the probability of a state as fixed, no matter which option we were evaluating.

We did, however, come across an exception to this pattern. Recall the lazy student: whether he studies is not independent of whether he passes, since studying raises the probability of passing. Even there, though, something important was still missing that will now become central. The state—passing or not passing the class—was not itself anyone's *choice*. It was just a fact about the world that happened to be sensitive, probabilistically, to what the student did. Nobody was choosing to fail the lazy student in order to satisfy their own preferences.

Games are what you get when you take that last step: when the “state” that determines how well your choice turns out is not a brute fact about the world, nor even a fact that is merely sensitive to your choice, but is itself the deliberate choice of another agent—an agent who, exactly like you, has preferences over the

possible outcomes and is choosing with an eye to satisfying them. This is what makes games distinctive, and it is why we call them *interactive* decisions.

Definition 10.1. Game: A *game* is a decision problem in which the relevant partition is made up not of states of nature, but of the possible choices of one or more other agents, each of whom has preferences over the possible outcomes and chooses with an eye to satisfying them. We will also call a game an **interactive decision**.

Because a game has more than one decision-maker, we need more than one utility function: one per player. For now we restrict ourselves to games with exactly two players, whom we will call *Row* and *Column*—names that will make sense once we get to the matrix representation below. For the sake of ease of exposition, we will also assume that Row is a man and Column is a woman. Row’s utility function will be written u_R , and Column’s u_C . Both take as arguments a pair consisting of a choice by Row and a choice by Column: $u_R(R_i, C_j)$ is how good the outcome is for Row, *by Row’s own lights*, when Row picks R_i and Column picks C_j ; and $u_C(R_i, C_j)$ is how good that very same outcome is for Column, *by Column’s lights*.

The rest of this chapter will make heavy use of two ways of depicting a game: as a matrix, and as a tree. We introduce the general schema for each before putting both to work on a simple example.

10.1.1 Matrices

The most compact way of representing a two-player game is as a matrix. Row’s available choices, $R_1, R_2, \dots, R_m$, label the rows; Column’s available choices, $C_1, C_2, \dots, C_n$, label the columns. Each cell of the resulting **payoff matrix** corresponds to one possible outcome—one for each combination of a choice by Row and a choice by Column—and so it must contain *two* utilities rather than one. We will adopt the following convention throughout: Column’s utility goes in the upper-right corner of the cell, and Row’s utility goes in the lower-left corner. Table 10.1 displays the general schema.

	C_1	C_2	$\dots$	C_n
R_1	$u_C(R_1, C_1)$ $u_R(R_1, C_1)$	$u_C(R_1, C_2)$ $u_R(R_1, C_2)$	$\dots$	$u_C(R_1, C_n)$ $u_R(R_1, C_n)$
R_2	$u_C(R_2, C_1)$ $u_R(R_2, C_1)$	$u_C(R_2, C_2)$ $u_R(R_2, C_2)$	$\dots$	$u_C(R_2, C_n)$ $u_R(R_2, C_n)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
R_m	$u_C(R_m, C_1)$ $u_R(R_m, C_1)$	$u_C(R_m, C_2)$ $u_R(R_m, C_2)$	$\dots$	$u_C(R_m, C_n)$ $u_R(R_m, C_n)$

Table 10.1: The general schema for a two-player payoff matrix. Row's choices label the rows, Column's choices label the columns, and each cell reports both players' utilities for that combination of choices: Column's utility in the upper-right corner, Row's in the lower-left.

Reading across a row tells you how well Row does against each of Column's possible choices; reading down a column tells you how well Column does against each of Row's possible choices. This is exactly the kind of information a single agent's decision matrix gave us in earlier chapters; the only novelty is that we now have two such matrices at once, layered into a single grid of cells.

10.1.2 Trees

Matrices are compact, but they hide the order, if any, in which the players choose, and they picture both players as choosing simultaneously and independently of one another. A second way of representing a game, its **extensive form** or *game tree*, makes the order of choice explicit. In a game tree, each point at which a player must choose is drawn as a circle, labeled with whichever player chooses there— R for Row, C for Column; the branches leading out of a node, each labeled with the choice it represents, are the options available to that player; and the tips of the branches are the possible outcomes, each labeled with a pair of utilities $(u_R(R_i, C_j), u_C(R_i, C_j))$, in that order. Figure 10.1 gives the schema for a two-player game in which Row chooses first, among $R_1, \dots, R_m$, and Column then chooses, among $C_1, \dots, C_n$, already knowing what Row chose.

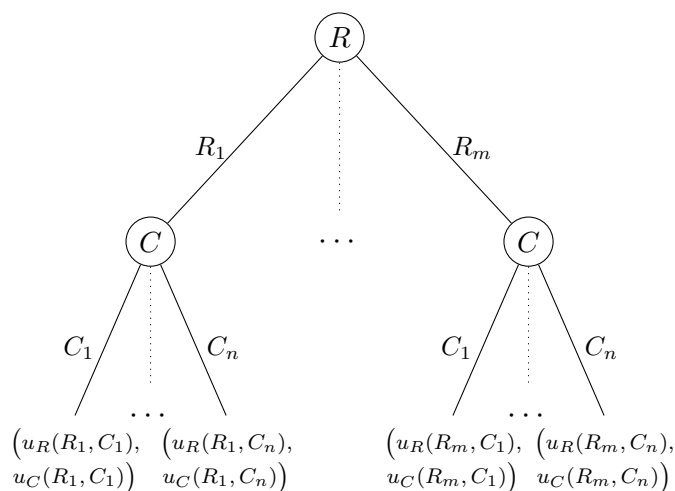


Figure 10.1: The general schema for a two-player game tree. Each choice node is drawn as a circle labeled with whichever player chooses there— R for Row, C for Column. Row chooses first, among $R_1, \dots, R_m$ (only R_1 and R_m are drawn; the dots stand for $R_2, \dots, R_{m-1}$). Column then chooses, among $C_1, \dots, C_n$. Each branch tip is labeled with the pair $(u_R(R_i, C_j), u_C(R_i, C_j))$: Row’s utility followed by Column’s.

Notice the qualification at the end of the last paragraph: in Figure 10.1, Column chooses *after* Row, branching out separately beneath each of Row’s choices, and so is implicitly pictured as knowing what Row chose before choosing herself. That is a fine way of depicting a genuinely sequential game. But many of the games that interest us—including the one we turn to next—are not sequential: both players choose at the same time, each in ignorance of what the other is doing. A tree can still represent such a game, but only if we add a device for marking that several nodes are, from the point of view of the player choosing at them, indistinguishable. We do this by joining those nodes with a dotted line; the resulting cluster is called an **information set**. A player choosing at an information set must choose the same way at every node in it, since she cannot tell which one she is actually at. We will have more to say about information sets later on; for now, the dotted line is just a reminder that a choice is being made in ignorance of another player’s choice.

10.1.3 Example: Rock-Paper-Scissors

Let us put both representations to work on a simple, familiar game. In *Rock-Paper-Scissors*, Row and Column simultaneously choose one of Rock, Paper, or Scissors. Rock beats Scissors, Scissors beats Paper, and Paper beats Rock; if both players

make the same choice, the game is a tie. The winner receives a utility of 1, the loser a utility of -1 , and both players receive 0 in the event of a tie.

Table 10.2 gives the matrix representation of the game.

	Rock	Paper	Scissors
Rock	0 0	1 -1	-1 1
Paper	-1 1	0 0	1 -1
Scissors	1 -1	-1 1	0 0

Table 10.2: Rock-Paper-Scissors in matrix form. As before, Column's utility appears in the upper-right corner of each cell and Row's in the lower-left. Row's choice of Rock against Column's choice of Paper, for instance, yields $u_R = -1$ and $u_C = 1$, since Paper beats Rock.

Figure 10.2 gives the tree representation of the very same game. Since Rock-Paper-Scissors is a simultaneous game, Column's three choice nodes—one beneath each of Row's choices—are joined by a dotted line, marking them as a single information set: Column must choose among Rock, Paper, and Scissors without having seen what Row chose.

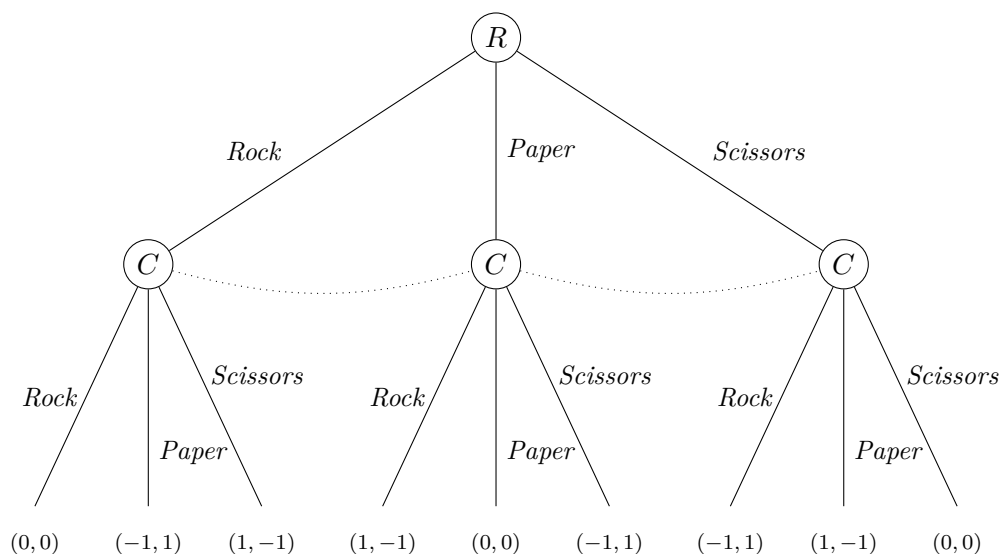


Figure 10.2: Rock-Paper-Scissors in tree form. Row chooses first, then Column chooses without knowing Row's choice—indicated by the dotted line joining Column's three choice nodes into a single information set. Each branch tip is labeled with the pair (u_R, u_C) of payoffs it yields.

The two representations agree, as they must, on every outcome: if Row plays Scissors and Column plays Paper, both the matrix and the tree tell us that Row gets a payoff of 1 and Column a payoff of -1 . Which representation is more convenient depends on the game and on what we want to do with it; we will use both freely throughout this chapter.

10.1.4 From trees to matrices, and back

Before we show how to translate one representation into the other, it helps to make explicit something that has so far been true without needing a name. In both Figure 10.1 and Figure 10.2, each player moves exactly once, so a row or a column of a matrix, and a first- or second-level branch of a tree, already amounts to a complete plan of action for that player. Once players may move more than once, this stops being true, and we need the more general notion of a *strategy*.

Definition 10.2. Strategy: A *strategy* for a player in a game is a complete plan of action for that player: a specification of what she would do at every point in the game—or, if several of her decision nodes belong to the same information set, at every information set of hers—at which it might be her turn to move.

In every game we have considered so far, each player has exactly one decision node (or one information set), so a strategy for Row just is one of $R_1, \dots, R_m$, and a strategy for Column just is one of $C_1, \dots, C_n$. The rows and columns of our matrices, and the branches of our trees, already are strategies. This is what lets the translation between the two representations be a matter of bookkeeping rather than something requiring us to introduce anything new. The bookkeeping goes as follows.

From a tree to a matrix. List Row’s strategies along the rows and Column’s strategies along the columns. For each pair of strategies—one for Row, one for Column—follow the tree from the root, letting each player move exactly as her strategy specifies; this traces a unique path down to a single leaf. Copy that leaf’s pair of utilities into the corresponding cell of the matrix. This is exactly how Table 10.2 was read off Figure 10.2: the strategy pair (Scissors, Paper), for instance, traces a path down the branch labeled *Scissors* and then the branch labeled *Paper*, landing on the leaf labeled $(1, -1)$ —which is exactly the entry in the Scissors row and Paper column of Table 10.2.

From a matrix to a tree. A matrix does not, by itself, say in what order, if any, the players moved; a tree, by contrast, always shows branches emerging from a root in some order. Turning a matrix into a tree therefore requires an arbitrary choice: pick one player to draw as moving “first” (it does not matter which), draw her strategies as the branches out of the root, and then, beneath *each* of those branches, draw a node for the other player, with her own strategies as its branches. If, as is usually the case, the matrix represents a genuinely simultaneous choice, join all of the second mover’s nodes into a single information set with a dotted line, exactly as we did for Column in Figure 10.2: this cancels out the false impression, created by the tree’s left-to-right layout, that the second mover already knows what the first one did. The leaves are then labeled directly from the matrix: the leaf at the end of the path through R_i and C_j gets whichever pair of utilities sits in row R_i , column C_j of the matrix.

Since Table 10.2 and Figure 10.2 depict the very same game, running either procedure on one of them should return (up to the arbitrary choice of who is drawn as moving first) the other. We encourage you to check this for yourself.

10.2 The assumptions of game theory

Before going any further, it is worth making explicit several assumptions that classical game theory takes for granted. None of them is beyond dispute—each has been challenged, by philosophers and economists alike—but we will adopt all four for the remainder of this chapter, since doing so is what lets us use matrices, trees, and the solution concepts to come in their standard form.

1. *Utilities are measured on interval scales.* Recall that an ordinal utility function only records a ranking of outcomes, whereas a cardinal, or interval-scale, utility function also records facts about the *magnitude* of the differences between them—and that this is exactly what maximizing expected utility requires (Norm 8.4). We will assume throughout that each player's utility function is at least interval-scale: u_R and u_C are each fixed only up to a positive linear transformation (Definition 8.1), so nothing follows from the bare fact that, say, $u_R(R_1, C_1) = 4$ while $u_C(R_1, C_1) = 2$. Comparisons of magnitude are licensed only *within* a single player's own scale—Row's 4 against one of Row's *other* numbers—never *across* players: classical game theory does not assume any way of comparing how much Row wants something to how much Column wants something.
2. *Players maximize expected utility.* Row and Column are both assumed to be maximizing expected utility (Norm 8.4), exactly as the single agents of earlier chapters did. It is worth being careful about what this assumption does, and does not, say. It does not say that a player always wants to *win*, in whatever sense her opponent would recognize as winning. A mother playing chess against her young son may well prefer that he win, and so assign a higher utility to that outcome than to her own victory. Game theory has no trouble representing this: it simply means that the mother and son are not playing the game that two rival grandmasters would be playing, even though the legal moves of chess are exactly the same in both cases. A game, recall, is individuated by the players' utilities over outcomes, not by the nominal activity they happen to be engaged in (Definition 10.1); change the utilities and, officially, you have changed the game, whatever the rules on the box say.
3. *All of this is common knowledge.* It is not enough that the rules, the payoffs, and each player's rationality merely happen to be known by both players; game theory assumes something considerably stronger.

Definition 10.3. Common knowledge: A proposition p is *common knowledge* among a group of agents if and only if every agent in the group knows p , every agent knows that every agent knows p , every agent knows that every agent knows that every agent knows p , and so on, without end.

This is a much stronger condition than what is called **mutual knowledge**: the mere fact that each member of a group individually knows p . Suppose I tell Row, in private, that Column is rational, and I separately tell Column, in private, that Row is rational. Then it is true that both know that the other is rational—the fact is mutually known. But neither of them knows that I told the other: for all Row knows, Column may think Row is irrational, and so Row cannot safely reason about what a rational Column will do *on the assumption that Column is reasoning about a rational Row*. Had I announced the very same fact *publicly*, with both of them present and aware of each other’s presence, it would have become common knowledge instead: each would know it, each would know that the other knows it, each would know that the other knows that they know it, and so on—because each can see that the other was there to hear the announcement, and can see that the other can see this too. It is this whole infinite hierarchy, not the mere fact that everyone happens to know, that licenses each player’s confidence in reasoning about what the other player, reasoning in turn about her, will do.

4. *We limit ourselves to two-person games.* For the rest of this chapter, every game we consider will have exactly two players, Row and Column, and no more. This is a substantive restriction—many interesting games have three or more players—but it keeps our matrices and trees to a manageable size, and the theory of games with more than two players is, in any case, considerably more complicated.

10.3 Strictly competitive two-player games

Look back at Table 10.2. Every one of its nine cells has the same curious feature: whatever number appears as Row’s utility, Column’s utility is exactly that number’s negative. A win for Row ($u_R = 1$) is paired with a loss for Column ($u_C = -1$); a loss for Row ($u_R = -1$) is paired with a win for Column ($u_C = 1$); a tie ($u_R = 0$) leaves them both exactly where they started ($u_C = 0$). Rock-Paper-Scissors is the simplest possible case of a wider phenomenon: games in which the players’

preferences are in exact opposition, so that to whatever extent one of them gains from an outcome, the other loses.

Definition 10.4. Strictly competitive: A two-player game is *strictly competitive* if the players' preferences are in exact opposition: for any two outcomes, or lotteries over outcomes, o and o' , Row prefers o to o' if and only if Column prefers o' to o .

Since each player's utility function is fixed only up to a positive linear transformation (our first assumption above), it makes no sense to ask whether Column's utility is "really" the negative of Row's. But in a strictly competitive game we are free to choose Column's utility function so that it is: letting $u(R_i, C_j)$ be Row's utility for a given outcome, we can take Column's utility for that same outcome to be $-u(R_i, C_j)$, so that $u_R(R_i, C_j) + u_C(R_i, C_j) = 0$ for every i and j . A game whose utilities are written in this way is called **zero-sum**. Every strictly competitive game can be written as a zero-sum game, and so we will use the two terms interchangeably from now on.

Since Column's utility is never independent information in a strictly competitive game—once we know Row's utility for an outcome, we know Column's for free, as its negative—we will simplify our matrices from here on by writing only Row's utility in each cell and leaving Column's implicit. Table 10.3 gives Rock-Paper-Scissors in this simplified notation. Compare it with Table 10.2 to confirm that no information has been lost: Column's utility in any cell can always be recovered just by flipping the sign of the number written there.

	Rock	Paper	Scissors
Rock	0	-1	1
Paper	1	0	-1
Scissors	-1	1	0

Table 10.3: Rock-Paper-Scissors in zero-sum form, with only Row's utility written in each cell. Column's utility in any cell is simply the negative of the number shown: Row's choice of Rock against Column's choice of Paper, for instance, yields $u_R = -1$, so $u_C = 1$.

10.4 Nonzero-sum games

Not all two-player games are zero-sum. Suppose, for instance, that Row and Column want to meet during this weekend, but cannot communicate. The options, which they arranged beforehand, are to go to the park or to the movies. They both

prefer meeting each other to not meeting, but Row prefers meeting at the park and Column meeting at the movies:

	Park	Movies
Park	1 2	0 0
Movies	0 0	2 1

Table 10.4: Park or Movies: the Game. Column's utility appears in the upper-right corner of each cell and Row's in the lower-left. Both players prefer meeting to not meeting, but Row prefers to meet at the park and Column at the movies.

10.5 Iterated deletion of dominated strategies

What are the norms for rational game players? One uncontroversial norm is that rational players do not play strongly dominated strategies. Remember our definition of a strongly dominated strategy from Chapter 7: option φ strongly dominates option ψ if and only if φ does better than ψ no matter which state obtains (Definition 7.8), that is, if and only if, for every state S_i ,

$$u(\varphi \wedge S_i) > u(\psi \wedge S_i).$$

In a game, the role of the states is played by the other player's strategies. From Row's point of view, what he does not control, and what together with his own choice settles the outcome, is which strategy Column plays; and the same goes, the other way around, for Column. So Row's strategy R_k strongly dominates his strategy R_i if and only if, for every strategy C_j of Column's,

$$u_R(R_k, C_j) > u_R(R_i, C_j),$$

and Column's strategy C_k strongly dominates her strategy C_j if and only if, for every strategy R_i of Row's,

$$u_C(R_i, C_k) > u_C(R_i, C_j).$$

In a zero-sum game written, as we have agreed, with Row's utilities only, the second condition has to be read with care. Since Column's utility is the negative of Row's, Column's strategy C_k strongly dominates C_j if and only if, for every R_i ,

$$u_R(R_i, C_k) < u_R(R_i, C_j).$$

That is, every number in column C_k is *lower* than the corresponding number in column C_j . Column, remember, wants the numbers in the matrix to be as low as possible.

The dominance norm (Norm 7.10) tells us that a rational player never plays a strongly dominated strategy. Consider, then, the following zero-sum game:

	C_1	C_2	C_3
R_1	0	2	6
R_2	4	3	5
R_3	9	1	7

Table 10.5: A zero-sum game, with only Row's utilities shown. Column's utility in each cell is the negative of the number shown, so Column prefers lower numbers.

None of Row's strategies is strongly dominated: R_1 does better than R_2 if Column plays C_3 ($6 > 5$), but worse if she plays C_1 ($0 < 4$); R_3 does better than both of the others against C_1 , but worse than both against C_2 . Column, however, has a strongly dominated strategy: every number in column C_3 is higher than the corresponding number in column C_2 ($6 > 2$, $5 > 3$, and $7 > 1$), so C_2 strongly dominates C_3 . A rational Column, therefore, will not play C_3 .

	C_1	C_2	C_3
R_1	0	2	6
R_2	4	3	5
R_3	9	1	7

Table 10.6: Step 1: Column deletes C_3 .

If the dominance norm were all we had to go on, that is where our analysis would stop. But recall our third assumption: the players' rationality, and the payoffs of the game, are common knowledge (Definition 10.3). Since Row knows that Column is rational, he knows that Column will not play C_3 , and so he can reason as if C_3 had been deleted from the game. And in the smaller game that results, R_1 is strongly dominated, by R_2 ($4 > 0$ and $3 > 2$). The only thing that protected R_1 was its good showing against C_3 , and C_3 is no longer a live possibility. So a rational Row who knows that Column is rational will not play R_1 .

	C_1	C_2	C_3
R_1	0	2	6
R_2	4	3	5
R_3	9	1	7

Table 10.7: Step 2: Row deletes R_1 .

The process does not stop there. Column knows that Row is rational and that Row knows that Column is rational, so she, too, can work out that Row will not play R_1 , and can reason as if both C_3 and R_1 had been deleted. In the game that is left, C_1 is strongly dominated by C_2 ($3 < 4$ and $1 < 9$): C_1 was protected only by the 0 that Row would get by playing R_1 , and Row will not play R_1 . So Column deletes C_1 as well.

	C_1	C_2	C_3
R_1	0	2	6
R_2	4	3	5
R_3	9	1	7

Table 10.8: Step 3: Column deletes C_1 .

Finally, Row can follow all of this reasoning, so he knows that Column will play C_2 ; and against C_2 , R_2 does better than R_3 ($3 > 1$). So Row deletes R_3 . The only strategies left standing are R_2 and C_2 , and the outcome of the game is that Row receives 3 (and Column, therefore, -3).

	C_1	C_2	C_3
R_1	0	2	6
R_2	4	3	5
R_3	9	1	7

Table 10.9: Step 4: Row deletes R_3 .

Notice how each step of this reasoning requires one more layer of knowledge than the one before. The first deletion requires only that Column be rational. The second requires, in addition, that Row know that Column is rational. The third requires that Column know that Row knows that Column is rational. And the fourth requires that Row know that Column knows that Row knows that Column is rational. In bigger games, the chain can get as long as you like, which is why game theory does not stop at any particular layer and simply assumes common knowledge, which delivers all of the layers at once.

Definition 10.5. Iterated deletion of strongly dominated strategies: Delete every strategy, of either player, that is strongly dominated. Then look at the smaller game that results and delete every strategy that is strongly dominated *there*; and so on, until no strongly dominated strategies remain.

Iterated deletion works in exactly the same way in games that are not zero-sum. The only difference is that we can no longer write just one number per cell, so we have to be careful to read each player's *own* utilities: Row's strategies are compared using the numbers in the lower-left corners, and Column's using the numbers in the upper-right corners. And since those upper-right numbers are now Column's own utilities, Column wants them to be as *high* as possible. Consider the game in Table 10.10.

	C_1	C_2	C_3
R_1	4 1	1 3	3 5
R_2	2 1	2 4	5 0
R_3	1 3	0 2	2 0

Table 10.10: A game that is not zero-sum. As usual, Column's utility appears in the upper-right corner of each cell and Row's in the lower-left.

This time it is Row who has a strongly dominated strategy to begin with: R_3 is strongly dominated by R_1 ($4 > 1$, $1 > 0$, and $3 > 2$), and also by R_2 ($2 > 1$, $2 > 0$, and $5 > 2$). None of Column's strategies is dominated yet: C_2 does better for Column than C_1 against R_1 and R_2 , but worse against R_3 ($2 < 3$). Once Column knows that Row will not play R_3 , however, C_1 is strongly dominated by C_2 ($3 > 1$ and $4 > 1$), so she deletes it. With C_1 gone, Row loses the one thing that made R_1 worth considering, its 4 against C_1 , and R_1 is now strongly dominated by R_2 ($2 > 1$ and $5 > 3$). Finally, with only R_2 left for Row, C_3 is strongly dominated by C_2 ($4 > 0$). The game is dominance solvable: Row plays R_2 , Column plays C_2 , and they receive 2 and 4 respectively. Figure 10.3 summarizes the process.

	C_1	C_2	C_3
R_1	1 4	3 1	5 3
R_2	1 2	4 2	0 5
R_3	3 1	2 0	0 2

Step 1: Row deletes R_3 .

	C_1	C_2	C_3
R_1	1 4	3 1	5 3
R_2	1 2	4 2	0 5
R_3	3 1	2 0	0 2

Step 3: Row deletes R_1 .

	C_1	C_2	C_3
R_1	1 4	3 1	5 3
R_2	1 2	4 2	0 5
R_3	3 1	2 0	0 2

Step 2: Column deletes C_1 .

	C_1	C_2	C_3
R_1	1 4	3 1	5 3
R_2	1 2	4 2	0 5
R_3	3 1	2 0	0 2

Step 4: Column deletes C_3 .

Figure 10.3: Iterated deletion of strongly dominated strategies in the game of Table 10.10. At each step, the cells belonging to deleted strategies are shaded; only (R_2, C_2) survives.

Something happens in this game that could never happen in a zero-sum game. Look at the outcome of R_1 and C_3 : Row would get 3 and Column would get 5, so *both* players would be better off there than at the solution of the game. Why, then, can they not get there? Because (R_1, C_3) is not stable. If Column plays C_3 , Row does better by switching to R_2 (5 rather than 3), which would leave Column with 0; and Column, who can see this coming, will not expose herself to that risk by playing C_3 . In a zero-sum game the question cannot even arise, since whatever is better for one player is worse for the other, so no outcome can be better for *both* players than another. Once we leave zero-sum games behind, by contrast, perfectly rational players, with common knowledge of each other's rationality, can end up at an outcome that both of them would gladly trade for another one. We will see a starker version of this phenomenon when we come to the Prisoner's Dilemma.

If, at the end of this process, each player is left with a single strategy, as happens in the two cases we examined in this section, we say that the game is **dominance solvable**, and we take the surviving pair of strategies to be its solution. Because we only ever delete *strongly* dominated strategies, the order in which we carry out the deletions makes no difference: whichever strongly dominated strategy we delete

first, we always end up with the same strategies at the end. The same is not true, as we will see, if we also delete weakly dominated strategies, where the order of deletion can affect the result.

Not every game is dominance solvable, however. In Rock-Paper-Scissors (Table 10.3), no strategy of either player is strongly dominated, or even weakly dominated: each of Row's strategies is the best possible against one of Column's strategies and the worst possible against another, and likewise for Column. So iterated deletion does not even get started, and we would need other tools to say anything about how such games should be played.

10.6 Nash equilibrium

Iterated deletion of dominated strategies is a powerful tool, but it only gets started if some player has a dominated strategy. Consider the zero-sum game in Table 10.11, written, as usual, with only Row's utilities.

	C_1	C_2	C_3
R_1	2	8	0
R_2	4	6	5
R_3	3	1	7

Table 10.11: A zero-sum game in which no strategy is dominated, with only Row's utilities shown.

No strategy here is dominated, not even weakly. Each of Row's strategies is his best option against one of Column's (R_1 against C_2 , R_2 against C_1 , and R_3 against C_3), and each of Column's strategies is her best option against one of Row's (C_1 against R_2 , C_2 against R_3 , and C_3 against R_1). So iterated deletion is of no help. Can the players nevertheless figure out what to do?

Since what is best for each player depends on what the other one does, it is natural for each of them to start by asking what the other will do. Suppose, for instance, that Row expects Column to play C_2 . Then his best reply is R_1 , which would get him 8. But Column knows that Row is rational, and she knows the payoffs, so she can work out that if Row expected her to play C_2 he would play R_1 ; and against R_1 , her best reply is not C_2 but C_3 , which holds Row down to 0. The expectation that Column will play C_2 undermines itself: if Row acted on it, Column, anticipating this, would not play C_2 after all. The same thing happens if Row expects Column to play C_3 : his best reply is R_3 , which would get him 7,

but Column's best reply to R_3 is C_2 , which holds him down to 1. In fact, these expectations chase each other round in a circle: C_2 leads to R_1 , which leads to C_3 , which leads to R_3 , which leads back to C_2 .

Things are different if Row expects Column to play C_1 . His best reply to C_1 is R_2 , and Column's best reply to R_2 is C_1 itself, since 4 is the lowest number in row R_2 . Here, and only here, the expectation confirms itself: if Row expects C_1 and therefore plays R_2 , then Column, who can anticipate this, has every reason to play C_1 , which is just what Row expected her to do. Running the same reasoning from Column's side singles out the very same pair. Once the players are at (R_2, C_1) , neither has any reason to move: Row gets 4, and switching would get him only 2 or 3; Column holds Row down to 4, and switching would let him have 6 or 5.

Notice that this solution is not particularly glamorous for Row. R_1 holds out the prospect of 8 and R_3 the prospect of 7, whereas the best R_2 can do is 6. But Row cannot count on getting 8 or 7: those outcomes require Column to play into his hands, and she has no reason to do so.

The same kind of reasoning works in games that are not zero-sum. In the game of Table 10.12, again, no strategy of either player is dominated.

	C_1	C_2	C_3
R_1	5 2	3 2	0 0
R_2	0 4	2 1	1 2
R_3	0 1	1 4	3 3

Table 10.12: A game that is not zero-sum, in which no strategy is dominated. Column's utility appears in the upper-right corner of each cell and Row's in the lower-left.

Suppose that Column expects Row to play R_1 . Her best reply would then be C_1 , which gets her 5 (rather than 3 or 0). But Row, anticipating this, would play R_2 against C_1 , since that gets him 4 (rather than 2 or 1). Against R_2 , however, Column would rather play C_2 (2 rather than 0 or 1); against C_2 , Row would rather play R_3 (4 rather than 2 or 1); and against R_3 , Column would rather play C_3 (3 rather than 0 or 1). At this point the chain of reasoning comes to rest, since Row's best reply to C_3 is R_3 itself (3 rather than 0 or 2). The pair (R_3, C_3) is the only one at which each player's strategy is a best reply to the other's, and wherever we start the chain of reasoning, it ends there. Each player gets 3.

Sometimes, however, this kind of reasoning does not single out a unique pair of strategies. Go back to the **Park or Movies** game (Table 10.4):

	Park	Movies
Park	1 2	0 0
Movies	0 0	2 1

Table 10.13: Park or Movies: the Game. Column's utility appears in the upper-right corner of each cell and Row's in the lower-left. Both players prefer meeting to not meeting, but Row prefers to meet at the park and Column at the movies.

If Row expects Column to go to the park, his best reply is to go to the park too (2 rather than 0); and if Column expects Row to go to the park, her best reply is to go there as well (1 rather than 0). So (Park, Park) confirms itself. But so does (Movies, Movies): if each expects the other at the movies, each does best by going there too (Row gets 1 rather than 0, and Column 2 rather than 0). The two remaining pairs do not confirm themselves: whoever expects the other at the park does best by going to the park, and whoever expects the other at the movies does best by going to the movies, so neither player has any reason to stick to a pair in which they miss each other. The reasoning that served the players so well in the previous two games thus leaves them, in this one, with two candidate solutions, and nothing in the game itself tells them which one to aim for. Worse, Row prefers one of them and Column the other. If each heads for his or her own favorite, they end up at (Park, Movies) and miss each other; and if each gallantly defers to the other, they end up at (Movies, Park) and miss each other again.

The self-confirming pairs of strategies we have been looking for are named after the mathematician John Nash, who introduced the idea in 1950. To state the definition precisely, we first need the notion of a best response.

Definition 10.6. Best response: A strategy R_i of Row's is a *best response* to a strategy C_j of Column's if and only if no strategy of Row's does better against C_j ; that is, if and only if, for every strategy R_k of Row's,

$$u_R(R_i, C_j) \geq u_R(R_k, C_j).$$

Likewise, C_j is a best response to R_i if and only if, for every strategy C_l of Column's,

$$u_C(R_i, C_j) \geq u_C(R_i, C_l).$$

Definition 10.7. Nash equilibrium: A pair of strategies (R_i, C_j) is a *Nash equilibrium* if and only if R_i is a best response to C_j and C_j is a best response to R_i .

Equivalently, a pair of strategies is a Nash equilibrium just in case neither player could do better by switching to another strategy while the other player sticks to hers. (R_2, C_1) is the only Nash equilibrium of Table 10.11, and (R_3, C_3) the only one of Table 10.12; Park or Movies has two, (Park, Park) and (Movies, Movies). You can also check that the solutions we found by iterated deletion in Table 10.5 and Table 10.10 (both of which happen to be (R_2, C_2)) are Nash equilibria. This is no accident: whenever iterated deletion of strongly dominated strategies leaves each player with a single strategy, the resulting pair is a Nash equilibrium, and it is the game's only one.

Things are different if we also delete weakly dominated strategies. Consider the game in Table 10.14.

	C_1	C_2
R_1	2 2	0 3
R_2	0 0	3 3

Table 10.14: A game in which deleting a weakly dominated strategy deletes a Nash equilibrium.

This game has two Nash equilibria. At (R_1, C_1) , Row's best response to C_1 is R_1 (2 rather than 0), and Column's best response to R_1 is C_1 (2 rather than 0). At (R_2, C_2) , Row's best response to C_2 is R_2 , since it gets him 3 and R_1 would get him no more than that, and Column's best response to R_2 is C_2 (3 rather than 0). Notice, moreover, that (R_2, C_2) is better for *both* players than (R_1, C_1) . But R_2 is weakly dominated by R_1 : against C_1 , R_1 does better ($2 > 0$), and against C_2 the two do equally well ($3 = 3$). If we delete R_2 on that basis, Column is left choosing between C_1 and C_2 against R_1 alone, and now C_2 is strongly dominated by C_1 ($2 > 0$). The procedure thus leaves us with (R_1, C_1) , and the equilibrium that both players prefer has been deleted along the way.

What has gone wrong? Deleting R_2 because it is weakly dominated makes sense only if Row takes seriously the possibility that Column will play C_1 , since that is the only case in which R_1 does better than R_2 . But if Row is confident that Column will play C_2 , as he is entitled to be at the equilibrium (R_2, C_2) , then

he loses nothing by playing R_2 . Strongly dominated strategies never raise this problem. A strongly dominated strategy does worse than some alternative against *everything* the other player might do, so it is never a best response to anything, and so it can never be part of a Nash equilibrium. Deleting strongly dominated strategies, however many times we do it, therefore never deletes an equilibrium. This is a second reason, in addition to the fact that the order of deletion can matter, to restrict iterated deletion (Definition 10.5) to strongly dominated strategies.

Not every game has a Nash equilibrium if we consider just the strategies depicted in the matrix corresponding to the game, however. Consider Rock-Paper-Scissors once more (Table 10.3), and go through each of the nine pairs of strategies:

	Rock	Paper	Scissors
Rock	0	-1	1
Paper	1	0	-1
Scissors	-1	1	0

Table 10.15: Rock-Paper-Scissors in zero-sum form.

In every one of them, at least one of the players would do better by switching. At (Rock, Rock), Row would rather switch to Paper (1 rather than 0); at (Rock, Paper), Row would rather switch to Scissors (1 rather than -1); at (Rock, Scissors), Column would rather switch to Paper (getting 1 rather than -1); and so on for the remaining six pairs. Whatever pair the players expect, one of them has a reason to depart from it, and the reasoning goes round in circles forever, just as it did for the strategies other than R_2 and C_1 in Table 10.11.

This is not quite the end of the story. Instead of simply choosing one of their strategies, players can also randomize: Row could, for instance, roll a die and play Rock if it lands 1 or 2, Paper if it lands 3 or 4, and Scissors if it lands 5 or 6. A way of choosing among one's strategies with given probabilities is called a **mixed strategy**; by contrast, the strategies we have been dealing with so far are called **pure strategies**, and what we have defined is, strictly speaking, a Nash equilibrium *in pure strategies*. Nash proved that every game with finitely many players and strategies has at least one Nash equilibrium once mixed strategies are allowed. In Rock-Paper-Scissors, for instance, the only equilibrium is the one in which each player plays each strategy with probability $1/3$: if Column does this, every strategy of Row's has the same expected utility, 0, so Row has nothing to gain by doing anything else, and the same goes for Column. We will not deal with mixed strategies in this book, however. From now on, "Nash equilibrium" will

always mean Nash equilibrium in pure strategies.

10.6.1 Saddle points

In zero-sum games, there is a simple, mechanical way of finding Nash equilibria. Recall that, in a zero-sum matrix written with Row's utilities only, Row wants the numbers to be as high as possible and Column wants them to be as low as possible. So Row's best response to a column is whichever row has the *largest* number in that column, and Column's best response to a row is whichever column has the *smallest* number in that row. A pair (R_i, C_j) is therefore a Nash equilibrium just in case the number in its cell is at once the largest in its column and the smallest in its row.

Definition 10.8. Saddle point: In a zero-sum game written with Row's utilities only, a cell is a *saddle point* if and only if its number is both the smallest number in its row and the largest number in its column.

The name comes from the shape of a horse's saddle: the middle of the seat is the lowest point as you move from the horse's head to its tail, but the highest point as you move from one of its sides to the other. In a zero-sum game, then, the Nash equilibria are exactly the saddle points.

To find the saddle points of a zero-sum game, write down the minimum of each row and the maximum of each column, as in Table 10.16. The minimum of a row is what Row can guarantee himself by playing it, whatever Column does; the largest of these row minima, which is called the **maximin** of the game, is therefore the most Row can guarantee himself. Choosing a row whose minimum is the maximin is just following the maximin norm of Norm 7.13, with Column's strategies playing the role of the states. Similarly, the maximum of a column is the most Row can get if Column plays it; the smallest of these column maxima, which is called the **minimax** of the game, is therefore the least that Column can guarantee to hold Row down to. (Since Column's utilities are the negatives of Row's, choosing a column whose maximum is the minimax is just Column following the maximin norm from her own point of view.)

	C_1	C_2	C_3	Row minimum
R_1	2	8	0	0
R_2	4	6	5	4*
R_3	3	1	7	1
Column maximum	4*	8	7	

Table 10.16: The game of Table 10.11, with the minimum of each row and the maximum of each column. The maximin and the minimax are starred; they are equal, and the cell where they meet, boxed, is a saddle point.

In Table 10.16, the maximin and the minimax are both 4, and the cell where the maximin row and the minimax column meet, (R_2, C_1) , is a saddle point: 4 is the smallest number in row R_2 and the largest in column C_1 . It is, of course, the equilibrium we reasoned our way to earlier. (You can check that the solution of Table 10.5, whose value is 3, is a saddle point as well.)

The maximin can never be larger than the minimax. Take any row R_i and any column C_j : the minimum of R_i is at most the number in their shared cell, which is in turn at most the maximum of C_j . So every row minimum is less than or equal to every column maximum, and in particular the largest row minimum is less than or equal to the smallest column maximum. When the two are equal, the game has a saddle point, at the intersection of any row that achieves the maximin with any column that achieves the minimax. When the maximin is strictly smaller than the minimax, the game has no saddle point, and so no Nash equilibrium. Rock-Paper-Scissors is a case in point: every row minimum is -1 and every column maximum is 1, so the maximin, -1 , falls short of the minimax, 1.

There is something striking about all this. When we first met the maximin norm, in decisions under ignorance, it could look excessively pessimistic: why assume that the worst state will obtain? In a zero-sum game, however, the “states” are chosen by an opponent whose interests are exactly opposed to yours and who can anticipate what you will do, so expecting the worst is just being realistic. And whenever a zero-sum game has a saddle point, each player’s cautious maximin choice is exactly an equilibrium choice: neither of them could do better by being less cautious.

Finally, zero-sum games never confront the players with the problem we saw in the Park or Movies case. A zero-sum game may have more than one saddle point, but all of them have the same value (namely, the maximin, which equals the minimax); and if (R_i, C_j) and (R_k, C_l) are both saddle points, then so are (R_i, C_l) and (R_k, C_j) . So each player can simply pick any of his or her equilibrium

strategies, without any need to coordinate with the other.

10.6.2 Finding pure equilibria by starring best responses

The saddle-point method only works for zero-sum games. Fortunately, there is an equally mechanical method that works for every game, zero-sum or not, and that follows the definition of a Nash equilibrium very closely. To see how it works, consider the game in Table 10.17, in which each player has five strategies.

	C_1	C_2	C_3	C_4	C_5
R_1	3 5	4 1	7 9	3 7	2 0
R_2	1 1	7 7	6 7	6 3	2 6
R_3	8 5	4 3	0 7	2 4	7 2
R_4	1 6	6 7	7 7	8 3	2 1
R_5	6 0	4 0	2 0	3 8	4 8

Table 10.17: A game with five strategies for each player.

We could, of course, go through all twenty-five cells one at a time, asking of each whether either player would do better by switching. But there is a quicker way, which consists of two passes through the matrix.

In the first pass, we go through the matrix column by column, and in each column we put a star next to Row's highest number, that is, next to the highest of the numbers in the lower-left corners. A star on Row's number in the cell of R_i and C_j thus records that R_i is a best response to C_j : against C_j , no other strategy of Row's does better. In column C_1 , for instance, Row's numbers are 5, 1, 5, 6, and 0, so we star the 6 in row R_4 . If several of Row's numbers are tied for highest in a column, we star all of them, since each of the corresponding strategies is a best response to that column. This happens in column C_2 , where Row's 7 in row R_2 and his 7 in row R_4 both get a star. The result of the first pass is shown in Table 10.18.

	C_1	C_2	C_3	C_4	C_5
R_1	3 5	4 1	7 9*	3 7	2 0
R_2	1 1	7 7*	6 7	6 3	2 6
R_3	8 5	4 3	0 7	2 4	7 2
R_4	1 6*	6 7*	7 7	8 3	2 1
R_5	6 0	4 0	2 0	3 8*	4 8*

Table 10.18: First pass: in each column, Row's highest number is starred, marking Row's best response to that column.

In the second pass, we go through the matrix row by row, and in each row we put a star next to Column's highest number, that is, next to the highest of the numbers in the upper-right corners. A star on Column's number in the cell of R_i and C_j records that C_j is a best response to R_i . In row R_1 , for instance, Column's numbers are 3, 4, 7, 3, and 2, so we star the 7 in column C_3 . Notice the switch: Row's stars are assigned column by column, because what we are looking for is Row's best response to each of Column's strategies, whereas Column's stars are assigned row by row, because what we are looking for is Column's best response to each of Row's strategies. The result of the second pass is shown in Table 10.19.

	C_1	C_2	C_3	C_4	C_5
R_1	3 5	4 1	7* 9*	3 7	2 0
R_2	1 1	7* 7*	6 7	6 3	2 6
R_3	8* 5	4 3	0 7	2 4	7 2
R_4	1 6*	6 7*	7 7	8* 3	2 1
R_5	6* 0	4 0	2 0	3 8*	4 8*

Table 10.19: Second pass: in each row, Column's highest number is starred as well, marking Column's best response to that row. The cells with two stars, (R_1, C_3) and (R_2, C_2) , are the Nash equilibria of the game.

Now look for the cells in which *both* numbers are starred. In Table 10.19 there

are two: (R_1, C_3) and (R_2, C_2) . These are the Nash equilibria of the game, and it is easy to see why a cell with two stars is guaranteed to be one. The star on Row's number in the cell of R_i and C_j tells us that R_i is a best response to C_j , and the star on Column's number tells us that C_j is a best response to R_i . But that is exactly what it takes, according to Definition 10.7, for (R_i, C_j) to be a Nash equilibrium. The reasoning also runs in the other direction: if (R_i, C_j) is a Nash equilibrium, then R_i is a best response to C_j , so Row's number in that cell gets a star in the first pass, and C_j is a best response to R_i , so Column's number gets a star in the second. So the method finds *all* the pure Nash equilibria of a game, and *only* them. In particular, if no cell ends up with two stars, the game has no Nash equilibrium in pure strategies.

The cells with just one star are instructive too. At (R_4, C_2) , Row is doing as well as he can given Column's choice, so his number is starred; but Column is not, since against R_4 she would rather switch to C_4 (8 rather than 6). At (R_4, C_4) the situation is reversed: Column's 8 is the best she can get against R_4 , but Row would rather switch to R_5 (8 rather than 3). A cell with only one star is one in which one of the players has a reason to switch, and that is enough to keep it from being an equilibrium. Notice, finally, that high numbers do not make an equilibrium: (R_4, C_3) gives both players 7, exactly as much as the equilibrium (R_2, C_2) , and yet it has no stars at all, since against C_3 Row would rather play R_1 (9 rather than 7), and against R_4 Column would rather play C_4 (8 rather than 7).

The saddle-point method for zero-sum games is just a special case of this one. In a zero-sum game written with Row's utilities only, Row's highest number in a column is the column's maximum, and Column's highest utility in a row corresponds to the row's *minimum*, since her utilities are the negatives of Row's. A cell gets both stars, then, just in case its number is the largest in its column and the smallest in its row: just in case it is a saddle point.

10.7 The Prisoner's Dilemma

10.7.1 The case

Row and Column have been arrested on suspicion of having committed a serious crime together. The police are holding them in separate cells, and they have no way of communicating with each other. The prosecutor has enough evidence to convict both of them of a minor offense, which carries a sentence of one year in prison, but not enough to convict either of them of the serious crime without a

confession. So the prosecutor visits each of them separately and makes each the same offer. If you confess and your partner stays silent, your testimony will be used against your partner, who will get ten years, and you will go free. If you both confess, you will each get five years. And if you both stay silent, you will each get one year, for the minor offense. Each prisoner knows that the other has received exactly the same offer, and each knows that the other knows this, and so on.

If we assume that each prisoner cares only about how long he or she spends in prison, and (as we did with money in Newcomb's problem) that utility is linear in years, we can represent each player's utility for an outcome as minus the number of years he or she spends in prison. The result is Table 10.20.

	Silent	Confess
Silent	-1 -1	0 -10
Confess	-10 0	-5 -5

Table 10.20: The Prisoner's Dilemma.

Consider the situation from Row's point of view. If Column stays silent, Row does better by confessing (he goes free, rather than spending a year in prison); and if Column confesses, Row again does better by confessing (five years rather than ten). So confessing strongly dominates staying silent for Row. A strategy that strongly dominates all of a player's other strategies is called a **dominant strategy**, and confessing is a dominant strategy for Row. The game is symmetric, so exactly the same reasoning shows that confessing is a dominant strategy for Column as well. By the dominance norm (Norm 7.10), then, neither player should stay silent: the solution of the game is that both confess, and each spends five years in prison. This is also the game's only Nash equilibrium, as you can check by starring best responses.

Notice that, unlike in the games we analyzed by iterated deletion, the conclusion that both prisoners confess does not depend on what the players know about each other. Each player's reason to confess is independent of what the other one does, and so of what the other one knows or believes: confessing is better whatever the other prisoner does. A perfectly rational prisoner will confess even if he has no idea whether his partner is rational, or what his partner thinks of him.

10.7.2 Pareto optimality

Many people find this result troubling. For there is another outcome, in which both prisoners stay silent, that is better for *both* of them: each would spend one year in prison rather than five. Two rational players, each doing what is clearly best for himself or herself, end up at an outcome that both of them consider worse than another outcome that was available to them. This is the “starker version” of the phenomenon we noticed in Table 10.10: there, too, the solution of the game was worse for both players than another outcome, but at least reaching that solution required iterated reasoning about what the other player knows. In the Prisoner’s Dilemma, each player’s own rationality is enough.

To describe the situation more precisely, we need the following notion, named after the economist Vilfredo Pareto.

Definition 10.9. Pareto optimality: An outcome x *Pareto dominates* an outcome y if and only if x is at least as good as y for every player and better than y for at least one player. An outcome is *Pareto optimal* if and only if no outcome Pareto dominates it.

In the Prisoner’s Dilemma, the outcome in which both prisoners stay silent Pareto dominates the outcome in which both confess, so the solution of the game is not Pareto optimal. Each of the other three outcomes is Pareto optimal: moving away from any of them would make at least one of the prisoners worse off. (Notice that in a zero-sum game every outcome is Pareto optimal, since any change that makes one player better off makes the other worse off. That is why a game like the Prisoner’s Dilemma could not be zero-sum.) The Prisoner’s Dilemma, then, is a game in which individually rational choices lead to a collectively irrational result, if by a collectively irrational result we mean one that is not Pareto optimal.

10.7.3 Two mistaken reactions

There are several natural reactions to the Prisoner’s Dilemma which, on reflection, turn out to be mistaken.

Changing the payoffs. One reaction is to point out that real criminals often do not find themselves in this situation, because there are often penalties for confessing. A prisoner who rats out his partner may face retaliation from the partner’s friends when he gets out, or lose face in his community, or simply be unable to live with himself. Suppose, for instance, that a confessor can expect

retaliation that is as bad, by his own lights, as six extra years in prison. Then the game the prisoners are playing is the one in Table 10.21.

	Silent	Confess
Silent	-1 -1	-6 -10
Confess	-10 -6	-11 -11

Table 10.21: The prisoners' game when confessing carries a penalty equivalent to six extra years in prison.

In this game, staying silent is a dominant strategy for both players ($-1 > -6$ and $-10 > -11$), and the solution is that both stay silent. But this is not a solution to the Prisoner's Dilemma. Recall that a game is individuated by the players' utilities (Definition 10.1): if confessing carries a penalty, then the prisoners' utilities are different, and so the game they are playing is not the Prisoner's Dilemma at all. The puzzle is about what rational players should do when they *are* in a Prisoner's Dilemma, and pointing out that people are sometimes in a different game does not answer it.

What the observation does show is something important about the design of social institutions. Prisoner's Dilemmas are not confined to prisons: whenever each member of a group does better by acting in a certain way whatever the others do, but everyone would be better off if no one acted that way (as with overfishing a shared lake, polluting a common river, or arms races between rival countries), the group faces a version of the dilemma. Laws, contracts, enforceable agreements, and social norms can all work by changing the payoffs, so that the game people actually play is no longer a Prisoner's Dilemma. There are good reasons, then, for a society to arrange things so that its members face as few Prisoner's Dilemmas as possible.

Blaming egoism. Another reaction is to blame the dilemma on the prisoners' selfishness: if only they cared about each other, the thought goes, they would both stay silent and both be better off. But egoism has nothing to do with why the dilemma arises. To see this, imagine two prisoners who are pure altruists: each cares *only* about how long the *other* one spends in prison, and not at all about his or her own sentence. And suppose that the prosecutor makes each of them a different offer from before. Each prisoner can either keep quiet or take the blame, claiming to have been the ringleader. If neither takes the blame, each gets three years. Taking the blame takes two years off your partner's sentence, but adds four

years to your own. So if Row takes the blame and Column does not, Row gets seven years and Column one; if both take the blame, each gets five years ($3 + 4 - 2$).

Since each prisoner's utility is now minus the number of years that the *other* prisoner spends in prison, the game they are playing is the one in Table 10.22.

	Keep quiet	Take the blame
Keep quiet	-3 -3	-1 -7
Take the blame	-7 -1	-5 -5

Table 10.22: The altruists' game. Each player's utility is minus the number of years the *other* player spends in prison.

This game has exactly the same structure as the Prisoner's Dilemma. Taking the blame is a dominant strategy for each player: whatever Column does, Row's taking the blame takes two years off Column's sentence, which is all Row cares about ($-1 > -3$ and $-5 > -7$), and the same goes for Column. So both take the blame, and each altruist ends up watching the other serve five years, when, had neither taken the blame, each would have had the satisfaction of seeing the other serve only three. The outcome in which both keep quiet Pareto dominates the outcome in which both take the blame, exactly as before. (Notice, by the way, that two *egoists* in this situation would face no dilemma at all: since taking the blame adds four years to your own sentence, each egoist would keep quiet, and they would end up at the Pareto optimal outcome.) What generates the dilemma, then, is not the content of the players' preferences, whether selfish or altruistic, but the structure of the game: each player has a dominant strategy, and the outcome in which both play their dominant strategies is Pareto dominated by another one.

10.7.4 The Prisoner's Dilemma and Newcomb's problem

The Prisoner's Dilemma bears a striking resemblance to Newcomb's problem, which we studied in the previous chapter. To bring the resemblance out, suppose that your partner in crime is your identical twin. The two of you grew up together, you think alike, and you have always made the same choices in the same circumstances. You are, let us say, 90% confident that, whatever you decide to do, your twin will decide the same. Your twin is in another cell, and nothing you do can have any causal influence on what your twin does. But your choice is nevertheless excellent evidence about your twin's.

From your point of view, the Prisoner's Dilemma with a twin is a decision problem in which the states (your twin stays silent, or your twin confesses) are probabilistically dependent on your choice, but not causally dependent on it: precisely the feature that made Newcomb's problem so controversial. Table 10.23 shows your utilities, together with the conditional probability of each state given each of your choices.

	Twin silent	Twin confesses
Silent	-1 0.9	-10 0.1
Confess	0 0.1	-5 0.9

Table 10.23: The Prisoner's Dilemma with a twin, as a decision problem for you, with your utility in the upper left of each cell and the conditional probability of each state given each choice in the lower right.

Confessing dominates staying silent, exactly as two-boxing dominates one-boxing in Newcomb's problem (Table 9.2): whatever your twin does, you spend less time in prison by confessing. But if we apply MCEU (Norm 8.6), staying silent comes out ahead:

$$CEU(\text{Silent}) = 0.9 \times (-1) + 0.1 \times (-10) = -1.9,$$

$$CEU(\text{Confess}) = 0.1 \times 0 + 0.9 \times (-5) = -4.5.$$

Staying silent plays the role of one-boxing, and confessing the role of two-boxing; your twin's choice plays the role of Predictor's prediction, and the reduction in your sentence that confessing secures whatever your twin does (one year if your twin stays silent, five if your twin confesses) plays the role of the extra \$1,000 in the transparent box. The arguments of the previous chapter carry over as well. The one-boxer's reasoning recommends staying silent: those who stay silent almost always end up with a short sentence, and those who confess almost always end up with a long one. The two-boxer's reasoning recommends confessing: your twin's choice is not caused by yours, so whatever your twin does, you will be better off confessing. As the philosopher David Lewis argued in a well-known paper, "Prisoners' Dilemma Is a Newcomb Problem" (1979), a Prisoner's Dilemma played with someone whose choice is a sufficiently reliable indicator of yours simply *is* a Newcomb problem, with your partner's choice playing the role of the prediction.

This also helps us see what the standard game-theoretic analysis of the Prisoner's Dilemma takes for granted. Classical game theory treats each player's choice as

probabilistically independent of the other's: learning what you will do tells you nothing about what the other player will do. When that is so, MCEU and MEU agree, and both agree with the dominance argument that you should confess. It is only when your choice is evidence about your partner's, as when your partner is your twin, that the controversy about Newcomb's problem reappears in the Prisoner's Dilemma.

10.8 Backwards Induction

10.8.1 Playing the Prisoner's Dilemma more than once

So far we have assumed that the prisoners play the Prisoner's Dilemma only once. What if they know that they will play it many times, against each other, learning after each round what the other player did? Would that change anything?

It may seem that it would. In a single play of the game, staying silent is simply a gift to the other player: whatever she does, you do worse by staying silent than by confessing. But when the game is played repeatedly, what you do in one round can influence what the other player does in later rounds. Staying silent in the early rounds can work as a *signal* that you are willing to cooperate, and a player who receives such a signal may reciprocate. Confessing, by contrast, invites retaliation. To think about these possibilities, we need to stop looking at each round in isolation, and look instead at the larger game, sometimes called a **repeated game**, whose moves are the rounds of the Prisoner's Dilemma. Each player's utility in the repeated game is the sum of his or her utilities in the individual rounds, and a strategy in the repeated game (Definition 10.2) is a complete plan specifying what to do in each round, depending on what has happened in the previous ones. Here is one such strategy, known as *tit for tat*: stay silent in the first round, and from then on, do whatever the other player did in the previous round. A player who plays tit for tat rewards silence with silence and punishes confession with confession. In a series of famous computer tournaments organized by the political scientist Robert Axelrod around 1980, in which strategies submitted by researchers from several disciplines played repeated Prisoner's Dilemmas against one another, tit for tat came out on top. If both players play tit for tat, both stay silent in every round, and both get the Pareto optimal outcome every time.

10.8.2 The backwards induction argument

There is, however, a problem with this reasoning, at least when the number of rounds is fixed in advance and both players know what it is. Suppose they know that they will play exactly one hundred rounds, and consider the last one. In round 100 there are no future rounds to influence: nothing a player does then can be rewarded or punished later. So round 100 is, in effect, a single play of the Prisoner's Dilemma, and in it confessing is a dominant strategy for each player. Both players, if rational, will confess in round 100, whatever has happened before.

Now consider round 99. Each player knows that the other is rational, and so knows that both of them will confess in round 100 no matter what happens in round 99. But then nothing done in round 99 can influence what happens later either: staying silent in round 99 cannot buy any cooperation in round 100, since there will be none to buy. So round 99, too, is in effect a single play of the Prisoner's Dilemma, and both players will confess in it. The same reasoning now applies to round 98: since both players know that both will confess in rounds 99 and 100 whatever happens in round 98, round 98 is in effect a single play of the game as well. Continuing in this way, round by round, we eventually reach round 1. The conclusion is that rational players will confess in every single round, and so will end up, round after round, at the outcome that each of them would gladly trade for mutual silence. The argument can be summarized as follows:

- (1) In the last round, both players will confess, since confessing is a dominant strategy in a single play of the game.
- (2) If both players will confess in every round after round k , whatever happens in round k , then round k is in effect a single play of the game, and so both players will confess in round k as well.

$\therefore$ Both players will confess in every round.

Arguments of this form, which start by working out what rational players would do at the very end of a game and then use that to work out what they would do at each earlier stage, are called backwards induction arguments.

Definition 10.10. Backwards induction: A method of solving games in which the players move in sequence, which first determines what a rational player would do at the last move of the game, then uses that to determine what a rational player would do at the second-to-last move, and so on back to the first move.

Notice that the argument depends crucially on the players' knowing when the game will end. If they do not know how many rounds there will be (if, for instance, after each round there is some chance that they will play again), then there is no last round from which the argument can start, and, provided the chance of playing again is high enough, strategies like tit for tat can once again make cooperation rational.

10.8.3 The centipede game

The repeated Prisoner's Dilemma is not the only game in which backwards induction leads to a distressing result. Consider the following game, known as the **centipede game** because of the shape of its tree. On a table there is a big pile of one-dollar bills: a million of them. Row moves first, and he has two options. He can go *down*: take two dollars from the pile and end the game. Or he can go *across*: take one dollar from the pile and pass the turn to Column, who then faces the same two options. The players keep alternating in this way, each taking a dollar and passing, until one of them goes down or the pile runs out. When there are only two dollars left, going down means taking both, while going across means taking one and leaving the last one to the other player, which ends the game. Since each move takes at least one dollar from the pile, the game lasts at most 999,999 moves, and the last of them belongs to Row. Figure 10.4 shows the tree of the game, assuming as usual that utilities are linear in money; at each leaf, Row's total winnings are listed first and Column's second.

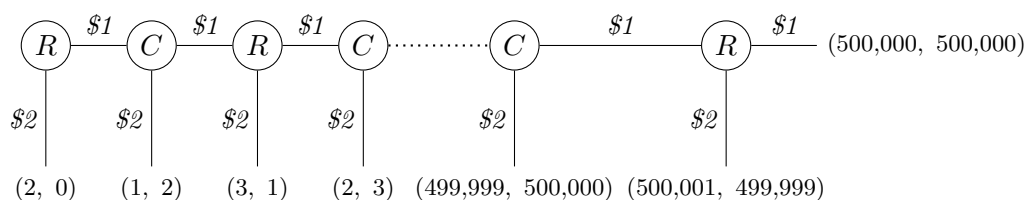


Figure 10.4: A centipede game with a pile of a million one-dollar bills. At each move, the player whose turn it is can go down, taking two dollars and ending the game, or go across, taking one dollar and passing the turn to the other player. The dotted line stands for moves 5 through 999,997; the last two nodes are moves 999,998 (Column's) and 999,999 (Row's). At each leaf, Row's total winnings are listed first and Column's second.

Unlike the Prisoner's Dilemma, this is a game of perfect information: the players move one at a time, and each knows, when it is his or her turn, everything that

has happened so far. And it is a game that rewards patience handsomely. If both players keep going across, they split the pile evenly, and each walks away with half a million dollars. If Row goes down at the very first move, he gets \$2 and Column gets nothing.

Now apply backwards induction. At the last move, move 999,999, there are two dollars left and it is Row's turn. Going down gets him both, for a total of \$500,001, whereas going across gets him one, for a total of \$500,000 (leaving the last dollar to Column); so, if he is rational, he goes down. At move 999,998, Column knows this, so she knows that if she goes across Row will go down, leaving her with \$499,999; going down now gets her \$500,000, so she goes down. At move 999,997, Row knows that if he goes across Column will go down, leaving him with \$499,999, whereas going down now gets him \$500,000; so he goes down. And so on: at every move, the player whose turn it is gets two dollars by going down, but only one by going across if the other player then goes down. At the very first move, then, Row goes down, and the game ends with Row getting \$2 and Column nothing, when each of them could have had half a million dollars. Backwards induction tells us that the game will end as soon as it begins.

10.8.4 The role of common knowledge

Both backwards induction arguments rely on the players' knowing a great deal about each other, and it is worth being precise about how much. Take the centipede game. Row's going down at the last move requires only that he be rational. Column's going down at the second-to-last move requires, in addition, that she know that Row is rational. Row's going down at the third-to-last move requires that he know that Column is rational and that he know that Column knows that he is rational. Each move further back requires one more layer, so that Row's going down at the very first move requires 999,998 layers: Row must know that Column knows that Row knows that Column knows ... (999,998 "knows" in all) ... that the player who moves last is rational. In general, in a centipede game with n moves the argument for going down at the first move needs $n - 1$ layers of knowledge, and in the hundred-round Prisoner's Dilemma the argument for confessing in round 1 needs 99. This is just the phenomenon we saw in our discussion of iterated deletion of dominated strategies: each step of the reasoning requires one more layer of knowledge than the one before. The assumption that rationality is common knowledge (Definition 10.3) delivers all the layers at once.

But now suppose that just *one* of those layers fails, however deep it is. In the

hundred-round Prisoner's Dilemma, for instance, suppose that the first ninety-eight layers hold, but the ninety-ninth does not. Then the backwards induction argument for confessing in round 1 breaks down: Row can no longer be sure that Column will confess in round 2 whatever happens in round 1, and so, for all Row knows, staying silent in round 1 may buy him Column's silence later on. And because the stakes in these games are so lopsided, it does not take much doubt to make cooperation rational. In the centipede game, suppose Row thinks that there is a probability p that Column is the kind of player who will keep going across as long as Row does, and a probability $1 - p$ that she will go down at the first opportunity. If Row goes down at once, he gets \$2. If he goes across, then with probability $1 - p$ Column goes down at the second move, leaving Row with \$1; and with probability p the game continues until Row goes down at the last move, leaving him with \$500,001. Going across has the higher expected utility as long as

$$500,001p + 1 \times (1 - p) > 2, \quad \text{that is, as long as } p > \frac{1}{500,000}.$$

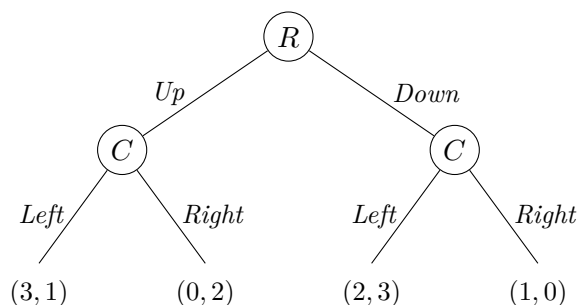
Any chance greater than one in half a million that Column will cooperate is enough to make going across rational for Row; and the same kind of reasoning applies at later moves, and to staying silent in the early rounds of a repeated Prisoner's Dilemma. So the distressing conclusions of the backwards induction arguments follow only if the players' rationality is known to a very great depth, and even a very high number of layers is not enough if it falls short of what the argument needs: ninety-eight layers will not do in a hundred-round Prisoner's Dilemma, and 999,997 will not do in our centipede game. Experiments fit this picture: when people play centipede games in the laboratory, very few of them go down at the first move; most go across for a while, and go down only as the end of the game approaches.

There is a final twist. Suppose that Row, despite the backwards induction argument, goes across at the first move, and Column finds herself at the second node. Column now knows something that, if rationality were common knowledge, could not have happened: Row has gone across. How should she reason from here? Should she conclude that Row is not rational after all, or that Row does not know that she is rational, or that Row is trying to signal something? Whatever she concludes, it seems that she can no longer take for granted the very assumptions on which the backwards induction argument relied. Philosophers and game theorists continue to disagree about how to reason in such situations, and so about whether common knowledge of rationality really does support the backwards induction

solution of games like the centipede.

Exercises

- 1. Strategies, trees, and matrices.** In the game below, Row first chooses Up or Down. Column sees what Row chose, and then chooses Left or Right. At each leaf, Row's utility is listed first and Column's second.



- Column has four strategies (Definition 10.2). List them. (Remember that a strategy for Column must say what she would do if Row played Up *and* what she would do if Row played Down.)
 - Write down the matrix of the game, with Row's two strategies as rows and Column's four strategies as columns.
 - Find all the pure Nash equilibria of the game by starring best responses.
 - Now suppose that Column does *not* see what Row chose before making her own choice. Draw the tree for this new game, write down its matrix, and find its pure Nash equilibria, if any. What difference does it make that Column can see Row's choice?
- 2. From matrices to trees.** Draw a tree for the Park or Movies game (Table 10.4), drawing Row as moving first. Then draw a second tree for the same game, this time drawing Column as moving first. Explain why the two trees represent the very same game.
- 3. Zero-sum games.** For each of the following games, say whether it is zero-sum. If it is, rewrite its matrix with Row's utilities only.

a)

	C_1	C_2
R_1	$\begin{matrix} -2 \\ 2 \end{matrix}$	$\begin{matrix} 1 \\ -1 \end{matrix}$
R_2	$\begin{matrix} 0 \\ 0 \end{matrix}$	$\begin{matrix} -3 \\ 3 \end{matrix}$

b)

	C_1	C_2
R_1	$\begin{matrix} 1 \\ 1 \end{matrix}$	$\begin{matrix} 2 \\ 0 \end{matrix}$
R_2	$\begin{matrix} 0 \\ 2 \end{matrix}$	$\begin{matrix} -1 \\ -1 \end{matrix}$

c)

	C_1	C_2
R_1	$\begin{matrix} 1 \\ 3 \end{matrix}$	$\begin{matrix} 2 \\ 2 \end{matrix}$
R_2	$\begin{matrix} 0 \\ 4 \end{matrix}$	$\begin{matrix} 3 \\ 1 \end{matrix}$

For game (c), show that a positive linear transformation of Column's utilities (Definition 8.1) turns it into a zero-sum game. Why is this transformation legitimate?

- 4. Iterated deletion.** Solve each of the following games by iterated deletion of strongly dominated strategies. At each step, say which strategy you delete and which strategy dominates it. Game (a) is zero-sum and is written, as usual, with Row's utilities only.

a)

	C_1	C_2	C_3
R_1	2	-1	-2
R_2	1	0	4
R_3	-3	-1	0

b)

	C_1	C_2	C_3
R_1	$\begin{matrix} 5 \\ 5 \end{matrix}$	$\begin{matrix} 0 \\ 0 \end{matrix}$	$\begin{matrix} 2 \\ 1 \end{matrix}$
R_2	$\begin{matrix} 2 \\ 3 \end{matrix}$	$\begin{matrix} 1 \\ 4 \end{matrix}$	$\begin{matrix} 5 \\ 0 \end{matrix}$
R_3	$\begin{matrix} 5 \\ 2 \end{matrix}$	$\begin{matrix} 0 \\ 3 \end{matrix}$	$\begin{matrix} 4 \\ 6 \end{matrix}$

For game (a), say how many layers of knowledge about the other player's rationality each step of the deletion requires.

- 5. Starring best responses.** Find all the pure Nash equilibria of each of the following games by starring best responses, or show that there are none. For game (c), also find all the Pareto optimal outcomes (Definition 10.9), and say whether the Nash equilibrium is one of them.

a)

	C_1	C_2	C_3
R_1	5 3	2 6	0 5
R_2	6 6	5 1	3 3
R_3	3 1	3 3	5 6

b)

	C_1	C_2	C_3
R_1	2 3	0 4	0 1
R_2	1 4	3 1	5 0
R_3	5 3	3 5	0 5

c)

	C_1	C_2	C_3	C_4
R_1	7 1	6 0	1 5	0 7
R_2	2 2	0 4	6 4	7 4
R_3	3 3	6 5	3 3	4 1
R_4	0 4	4 0	5 2	6 4

- 6. Saddle points.** Each of the following zero-sum games is written with Row's utilities only. For each one, write down the minimum of each row and the maximum of each column, find the maximin and the minimax, and find all the saddle points, if any. For the game with more than one saddle point, check that they all have the same value.

a)

	C_1	C_2	C_3	C_4
R_1	2	1	2	0
R_2	4	-4	0	0
R_3	1	3	2	1

b)

	C_1	C_2	C_3
R_1	4	0	4
R_2	1	-1	6
R_3	3	8	-3

c)

	C_1	C_2	C_3
R_1	1	1	1
R_2	-3	0	4
R_3	-2	-3	6

7. Weak dominance. Consider the following game.

	C_1	C_2	C_3
R_1	1 1	3 4	3 2
R_2	4 1	1 2	3 3

- Find all the pure Nash equilibria of the game.
- Show that C_2 is weakly, but not strongly, dominated.
- Delete C_2 , and keep deleting weakly dominated strategies until no more can be deleted. What is left? Which Nash equilibrium of the original game has been lost along the way?

8. Prisoner's Dilemmas and their relatives. For each of the following games, answer these questions: Does either player have a dominant strategy? What are the pure Nash equilibria? Which outcomes are Pareto optimal? Is the game a Prisoner's Dilemma?

a)

	C_1	C_2
R_1	3 3	5 0
R_2	0 5	1 1

b)

	C_1	C_2
R_1	4 4	3 0
R_2	0 3	2 2

c)

	C_1	C_2
R_1	4 4	3 2
R_2	2 3	1 1

9. Playing with a twin. You are playing the Prisoner's Dilemma of Table 10.20 against your twin, and you are 80% confident that, whatever you decide to do, your twin will decide the same.

- Compute the conditional expected utility of staying silent and of confessing. Which option does MCEU recommend? Which option does dominance reasoning recommend?
- Let r be your confidence that your twin will decide as you do. For which values of r does MCEU recommend staying silent?

10. A short centipede. Consider the centipede game of Figure 10.4, but with a pile of only six one-dollar bills.

- Draw the tree of the game, listing each player's total winnings at each leaf.
- Use backwards induction to find out what rational players would do. What are the players' winnings?
- How many layers of knowledge of rationality does the backwards induction argument require for Row's first move?
- Suppose Row thinks that there is a probability p that Column is the kind of player who will keep going across as long as Row does, and a probability $1 - p$ that she will go down at her first opportunity. For which values of p is going across at the first move better for Row than going down?

11. Conceptual questions.

- a) A teacher tells each of her students, in private, that tomorrow's exam has been cancelled. Is the cancellation mutually known among the students? Is it common knowledge (Definition 10.3)? What could the teacher do to make it common knowledge?
- b) A mother plays chess against her young son, and wants him to win. Two grandmasters play chess against each other, and each wants to win. Explain why, according to game theory, these are not the same game, even though the rules of chess are the same in both cases.
- c) Explain why the backwards induction argument for confessing in every round of a repeated Prisoner's Dilemma cannot get started if the players do not know how many rounds they will play.
- d) Describe a situation from everyday life, other than those mentioned in the chapter, that has the structure of a Prisoner's Dilemma. Then describe a law, contract, or social norm that could change the payoffs so that the people involved are no longer in a Prisoner's Dilemma.

Appendix: Solutions to the Exercises

Solutions are numbered as the exercises are (headings inside an exercise list do not consume a number). Where an exercise admits several correct answers, one is given.

Chapter 1: Language and Logic

1. For example: “Leona is a cat,” “Leona is a member of the species *felis catus*,” and “Leona est une chatte.” Three different sentences, one proposition.
2. Yes: a rhetorical question (“Who doesn’t love a free lunch?”) is interrogative in form, but it is used to assert something, so it can express a proposition.
3. Each of the n atomic sentences is independently assigned T or F. Adding one more atomic sentence doubles the number of assignments (each old assignment extends in two ways), so there are 2^n rows.
4. The table is fine for the particular sentences A, B . A *definition* of a connective must cover arbitrary wffs, so it has to be stated with metavariables φ, ψ rather than A, B .
5. $\varphi \wedge \psi \equiv \neg(\neg\varphi \vee \neg\psi)$ and $\varphi \vee \psi \equiv \neg(\neg\varphi \wedge \neg\psi)$ (De Morgan).
6. With NAND ($\uparrow$): $\neg\varphi \equiv \varphi \uparrow \varphi$; $\varphi \wedge \psi \equiv (\varphi \uparrow \psi) \uparrow (\varphi \uparrow \psi)$; $\varphi \vee \psi \equiv (\varphi \uparrow \varphi) \uparrow (\psi \uparrow \psi)$. With NOR ($\downarrow$): $\neg\varphi \equiv \varphi \downarrow \varphi$; $\varphi \vee \psi \equiv (\varphi \downarrow \psi) \downarrow (\varphi \downarrow \psi)$; $\varphi \wedge \psi \equiv (\varphi \downarrow \varphi) \downarrow (\psi \downarrow \psi)$.
7. Columns are listed in the row order TT, TF, FT, FF (for (i), rows TTT, TTF, TFT, TFF, FTT, FTF, FFT, FFF).

Wff?	Truth values	Type
(a) no (unbalanced parentheses)		
(b) yes, using the conventions only	TTTT	tautology
(c) no		
(d) yes, using the conventions only	TTFF	contingency ($\equiv A$)
(e) no, either way (no unique main connective)		
(f) yes, using the conventions only	TTTT	tautology
(g) yes, using the conventions only	TTTT	tautology (not a contradiction)
(h) yes, using the conventions only	TFFT	contingency
(i) yes, using the conventions only	TTTTTFFT	contingency (false when $A=F, B=T, C=F$)
(j) never a wff (ends in $\wedge$)		

Chapter 2: Translations

1. Exclusive or: $(\varphi \vee \psi) \wedge \neg(\varphi \wedge \psi)$, equivalently $\neg(\varphi \leftrightarrow \psi)$.
2. $(N \wedge \neg W) \rightarrow B$. (N : the weather is nice; W : Alyssa has work; B : Alyssa plays basketball.)
3. $\neg R \rightarrow \neg L$ (equivalently $\neg L \vee R$). (R : the strawberries are ripe; L : Bob likes the strawberries.)
4. $G \leftrightarrow (S \wedge E)$. (G : Charly graduates; S : she has satisfied her curriculum requirements; E : she has enough credits.)
5. $(S \vee W) \rightarrow F$. (S : the stairs are slippery; W : the stairs are wet; F : Deion falls.)
6. $(T \wedge M) \wedge \neg B$. (T : Erin likes to watch TV shows; M : she likes to watch movies; B : she likes to play basketball.)
7. $\neg P \leftrightarrow (\neg M \wedge \neg D)$. (P : Francis goes to the party; M : his mother gives permission; D : his father gives permission.)
8. $(G \rightarrow H) \wedge (\neg G \rightarrow V)$. (G : the weather is good; H : Gianna goes hiking; V : she plays video games.)
9. $T \rightarrow (S \vee N)$. (T : Hadrian gets a trophy; S : his soccer team wins its championship; N : his tennis team wins its championship.)
10. $(W \vee L) \rightarrow S$. (W : Ida works hard; L : she is lucky; S : she is successful.)
11. $((I \vee T) \wedge \neg M) \rightarrow K$. (I : Joel eats ice cream; T : he eats tomatoes; M : he takes his medication; K : he gets sick.) The sentence is ambiguous; on the other

reading it is $(I \vee (T \wedge \neg M)) \rightarrow K$.

Chapter 3: Arguments and Validity

1. Valid.
2. Valid.
3. Invalid: $F=F$, $G=T$ makes the premises true and the conclusion false.
4. Valid.
5. Invalid: $P=F$, $Q=T$.
6. Valid.
7. Valid.
8. Invalid: $P=F$, $Q=F$.
9. $I \rightarrow M$, $H \rightarrow \neg M \therefore I \rightarrow \neg H$ (I : Jasmine is interested in math; M : she takes a math class; H : she is a history major). Valid: if I and H were both true, M and $\neg M$ would both be true.
10. $D \rightarrow J$, $D \rightarrow P \therefore P \rightarrow J$ (D : Curtis is a great dancer; J : he gets into Juilliard; P : he practices). Invalid: $D=F$, $J=F$, $P=T$.
11. $A \rightarrow \neg E$, $E \therefore \neg A$ (A : Luke paid attention; E : his relationship ended). Valid (modus tollens).
12. $(C \vee H) \rightarrow G \therefore G \rightarrow H$ (C : Aliyah is clever; H : she is hard working; G : she gets good grades). Invalid: $C=T$, $H=F$, $G=T$.
13. $W \rightarrow F$, $\neg W \therefore \neg F$ (W : Ryan Gosling won an Academy Award; F : he is famous). Invalid (denying the antecedent): $W=F$, $F=T$.

Chapter 4: Measures and Stochastic Truth-Tables

1. $P(A \leftrightarrow B) = 2/5$.
2. $P(C \rightarrow B) = 1$.
3. $P(A \wedge C) = 1/10$.
4. $P(\neg A \wedge C) = 1/10$.
5. The column $P(\cdot \mid B \leftrightarrow C)$ gives, for rows TTT, TTF, TFT, TFF, FTT, FTF, FFT, FFF: $1/6$, 0 , 0 , $1/2$, $1/6$, 0 , 0 , $1/6$ (each old value divided by $P(B \leftrightarrow C) = 3/5$, and 0 outside the $B \leftrightarrow C$ rows).

6. Yes, it is possible. One distribution that works, for rows TTT, ..., FFF in the same order: $1/5$, 0 , 0 , $1/4$, $1/5$, $1/10$, 0 , $1/4$. It sums to 1 and satisfies all four constraints.

Chapter 5: Axioms, Definitions, and Theorems

1. (a) independent; (b) not independent; (c) independent; (d) not independent.
2. (a) Not the fallacy: drawing without replacement, the probability that the next card is an ace is $4/26$, not $4/52$, so the earlier draws really are relevant. (b) The fallacy. (c) The fallacy (in its reverse form). (d) Not the fallacy.
3. $\varphi \vee \psi \equiv (\varphi \wedge \neg\psi) \vee \psi$, and the disjuncts are mutually exclusive, so $P(\varphi \vee \psi) = P(\varphi \wedge \neg\psi) + P(\psi)$. Also $\varphi \equiv (\varphi \wedge \psi) \vee (\varphi \wedge \neg\psi)$, again mutually exclusive, so $P(\varphi \wedge \neg\psi) = P(\varphi) - P(\varphi \wedge \psi)$. Substituting,

$$P(\varphi \vee \psi) = P(\varphi) + P(\psi) - P(\varphi \wedge \psi).$$

4. By monotonicity, $P(\varphi \wedge \psi) \leq P(\varphi) = 0$, so $P(\varphi \wedge \psi) = 0$. Hence $P(\varphi | \psi) = 0 = P(\varphi)$ whenever $P(\psi) > 0$.
5. (a) independent ($1/36 = \frac{1}{6} \cdot \frac{1}{6}$); (b) not independent ($1/12$ vs. $1/18$); (c) independent ($1/12$); (d) not independent (0 vs. $1/144$); (e) independent ($1/18$).
6. (a) $3/4$; (b) $5/9$; (c) 1 ; (d) $11/12$; (e) $1/8$; (f) $9/14$; (g) $8/15$; (h) 0 ; (i) 0 ; (j) $3/7$; (k) $1/3$; (l) $2/3$; (m) 1 ; (n) $3/4$; (o) yes; (p) no (e.g. $P(H | G) = 2/3 \neq 1/3 = P(H)$); (q) $9/10$; (r) $1/2$.
7. (a) $2/5$; (b) $2/25$; (c) $1/5$; (d) $1/5$; (e) $2/5$.
8. (a) $3/5$ (this is $P(M)$); (b) $3/5$ ($P(J | M)$); (c) $4/5$ ($P(J | L)$); (d) 1 ($P(K \vee M)$); (e) $2/5$ ($P(J \wedge L)$). Along the way, $P(L) = 1/2$.
9. (a) No. (b) $3/4$.

Chapter 6: Bayesianism

1–3 (plague). $7/26 \approx 27\%$; 1 ; $1/58 \approx 1.7\%$.

4–6 (drunk). $15/22 \approx 68\%$; $57/71 \approx 80\%$; $5/9 \approx 56\%$.

7–8 (taxi). $2/3$; $8/9$.

9–10 (art). (9) $1/5$: the likelihood ratio is 1, so the evidence leaves the prior unchanged. (10) 1 : the museum's displaying the work is the declaration of authenticity, so $P(M | \neg A') = 0$ and the posterior is 1.

Chapter 7: Decisions under Ignorance

1. Multiplying by -1 reverses the order of the outcomes, and multiplying by 0 makes every outcome tie, so neither transformation preserves the ordinal ranking. Neither is acceptable.
2. (a) Maximin: the row minima are $-3, 2, -10$, so choose A_2 . (b) The acts tie on their minima, so apply leximin: the sorted utility vectors are $(0, 1, 1, 3)$, $(0, 1, 2, 4)$, $(0, 0, 1, 3)$, and A_2 is lexicographically best.
3. (a) Regrets: $A_1 = (3, 21, 7)$, max 21; $A_2 = (5, 0, 0)$, max 5; $A_3 = (0, 26, 12)$, max 26. Minimax regret chooses A_2 . (b) Regrets: $A_1 = (0, 0, 4)$, max 4; $A_2 = (8, 12, 0)$, max 12; $A_3 = (1, 15, 1)$, max 15. Minimax regret chooses A_1 .

Chapter 8: Risk and Expected Utility

Ex.	Norm	Values	Choice
1	MCEU	$T : 8.6$ vs. $\neg T : 7$	practice
2	MEU	$A : 5.3$ vs. $B : 4.55$	A
3	MEU	red 2.8 vs. black 1.76	red
4	MCEU	$E : 7.4$ vs. $\neg E : 6.6$	exercise
5	MEU	$U : 3.8$ vs. $\neg U : 4.0$	leave the umbrella
6	MCEU	$C : 8.84$ vs. $\neg C : 8.5$	put the phone away
7	MCEU	$V : 8.75$ vs. $\neg V : 8.0$	vaccinate
8	MCEU	$H : 6.4$ vs. $\neg H : 5.7$	cold-call

Chapter 9: Newcomb's Problem

1. With reliability $r = 0.7$: $\text{CEU}(\text{one-box}) = 700,000$ and $\text{CEU}(\text{two-box}) = 301,000$. MCEU recommends one-boxing.
2. With the \$10M prize: 9,000,000 vs. 1,001,000. One-box.
3. $\text{CEU}(O) = 1,000,000r$ and $\text{CEU}(\neg O) = 1,001,000 - 1,000,000r$. They tie at $r = 1001/2000 = 0.5005$; for r above this, one-box, and below it, two-box.
4. The tie occurs at $x = 800,000$.

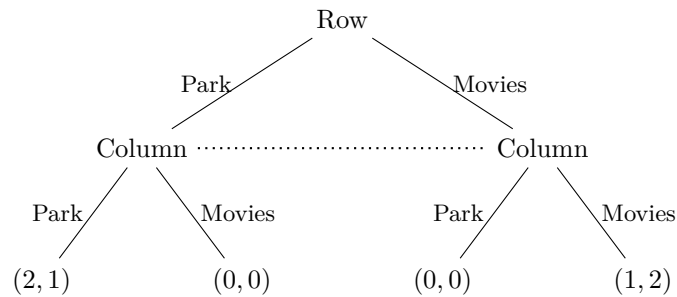
Chapter 10: Game Theory

1. Column's strategies are LL, LR, RL, RR, where each letter is the response to Up and then to Down. Pairs are (Row, Column):

	LL	LR	RL	RR
Up	(3, 1)	(3, 1)	(0, 2)	(0, 2)
Down	(2, 3)	(1, 0)	(2, 3)	(1, 0)

The unique Nash equilibrium is (Down, RL), with outcome (2, 3). In the simultaneous version (leaves (3, 1), (0, 2), (2, 3), (1, 0)) there is no pure-strategy equilibrium. So Column's observing Row's move makes the difference: with observation the game has exactly one equilibrium, (2, 3); without it, none.

2. Two trees, with leaves (u_R, u_C) equal to (2, 1), (0, 0), (0, 0), (1, 2), one for each player moving first. In the tree where Row moves first, Column's two decision nodes are joined by a dotted line (one information set), since Column cannot see Row's move:



For the other tree, Column moves first and Row's nodes are joined, with the leaves written as (u_R, u_C) in the same pattern: Park–Park (2, 1), Park–Movies (0, 0), Movies–Park (0, 0), Movies–Movies (1, 2). The matrix does not encode the order of play; the trees do, and they also show what each player can observe.

3. (a) Zero-sum; the Row-only matrix is $\begin{pmatrix} 2 & -1 \\ 0 & 3 \end{pmatrix}$. (b) Not zero-sum. (c) Constant-sum with total 4. Utilities are fixed only up to a positive linear transformation, so subtracting 4 from Column's utilities ($a = 1$, $b = -4$) changes neither choices nor equilibria; it gives the Row-only matrix $\begin{pmatrix} 3 & 2 \\ 4 & 1 \end{pmatrix}$.
4. (a) Delete R3 (strictly dominated by R2), then C1 (by C2), then R1 (by R2), then C3 (by C2). What remains is (R2, C2), with value 0. Dominance is judged in the reduced game, so the deletions need 0, 1, 2, 3 prior layers. (b) Delete C2 (by C1), R2 (by R1), C3 (by C1), R3 (by R1). What remains is (R1, C1) = (5, 5).

5. (a) Two equilibria: $(R2, C1) = (6, 6)$ and $(R3, C3) = (6, 5)$. (b) None. (c) A unique equilibrium, $(R3, C2) = (5, 6)$. The Pareto optimal outcomes in (c) are $(R1, C4) = (7, 0)$, $(R2, C4) = (4, 7)$ and $(R3, C2) = (5, 6)$, so the equilibrium is Pareto optimal.
6. (a) Maximin = minimax = 1; one saddle point, $(R3, C4)$. (b) Maximin $0 <$ minimax 4; no saddle point. (c) Maximin = minimax = 1; saddle points $(R1, C1)$ and $(R1, C2)$, both of value 1.
7. The Nash equilibria are $(R1, C2) = (4, 3)$ and $(R2, C1) = (1, 4)$. Column's C3 weakly (not strongly) dominates C2. Deleting C2, then R1 (now weakly dominated by R2), then C3 leaves only $(R2, C1)$. The equilibrium lost by iterated deletion of weakly dominated strategies is $(R1, C2)$.
8. (a) Prisoner's dilemma: R2 and C2 are dominant, the unique equilibrium is $(1, 1)$, and every outcome except $(1, 1)$ is Pareto optimal. (b) Stag hunt: no dominant strategies; equilibria $(R1, C1)$ and $(R2, C2)$; only $(4, 4)$ is Pareto optimal. Not a prisoner's dilemma. (c) R1 and C1 are dominant; the equilibrium is $(4, 4)$, the only Pareto optimal outcome. Not a prisoner's dilemma.
9. With $r = 0.8$: $CEU(\text{Silent}) = -2.8$ and $CEU(\text{Confess}) = -4$, so MCEU says to stay silent while dominance says to confess. In general $CEU(S) = -10 + 9r$ and $CEU(C) = -5r$, so Silent is recommended exactly when $r > 5/7$.
10. (a) The tree alternates Row, Column, Row, Column, Row; at each node the mover chooses *down* (ending the game) or *across*. The leaves reached by *down* are, in order, $(2, 0)$, $(1, 2)$, $(3, 1)$, $(2, 3)$, $(4, 2)$, and the final *across* gives $(3, 3)$. (b) By backwards induction, Row goes down at the first move, giving $(2, 0)$. (c) Four layers of knowledge are needed. (d) Going across is better iff $1 + 3p > 2$, i.e. $p > 1/3$.
11. (a) Mutual knowledge, yes; common knowledge, no; the announcement should be public. (b) Games are individuated by their utilities, not by their outward description: the mother's goal is that her son win, so her utilities differ from what the rules alone suggest and the game is a different one. (c) Backwards induction starts from the last move; if there is no last round (or the players do not know when it is), there is no starting point for the induction, and cooperation can be sustained. (d) An everyday prisoner's dilemma together with a law, contract or norm that changes the payoffs, so that the strategy that was dominated becomes unattractive.

Appendix: Formula Sheet

- Def. 4.1 **MUTUALLY EXCLUSIVE.** Two propositions φ and ψ are mutually exclusive iff there is no row of their truth-table in which both are true. (So $\varphi \wedge \psi$ is a contradiction, and $P(\varphi \wedge \psi) = 0$.)
- Def. 4.2 **CONDITIONAL PROBABILITY.** For any φ and ψ , if $P(\psi) \neq 0$, then $P(\varphi | \psi) = P(\varphi \wedge \psi) / P(\psi)$.
- Def. 5.2 **INDEPENDENCE.** φ is independent of ψ iff $P(\psi) \neq 0$ and $P(\varphi | \psi) = P(\varphi)$. φ is *dependent* on ψ iff $P(\psi) \neq 0$ and $P(\varphi | \psi) \neq P(\varphi)$. φ and ψ are independent *of each other* iff each is independent of the other.
- Ch. 6 **CONDITIONAL INDEPENDENCE.** φ and χ are independent conditional on ψ iff $P(\varphi \wedge \chi | \psi) = P(\varphi | \psi) \times P(\chi | \psi)$.
- Axiom 1 **NON-NEGATIVITY.** For any φ , $P(\varphi) \geq 0$.
- Axiom 2 **NORMALITY.** If τ is a tautology, then $P(\tau) = 1$.
- Axiom 3 **ADDITIVITY.** If φ and ψ are mutually exclusive, then $P(\varphi \vee \psi) = P(\varphi) + P(\psi)$.
- Thm. 5.1 **CONTRADICTION.** If φ is a contradiction, then $P(\varphi) = 0$.
- Thm. 5.5 **COMPLEMENT.** For any φ , $P(\varphi) + P(\neg\varphi) = 1$.
- Thm. 5.6 **EQUIVALENCE.** If φ and ψ are tautologically equivalent, then $P(\varphi) = P(\psi)$.
- Thm. 5.9 **EXTENDED ADDITIVITY.** If $\varphi_1, \dots, \varphi_n$ are pairwise mutually exclusive, then $P(\varphi_1 \vee \dots \vee \varphi_n) = P(\varphi_1) + \dots + P(\varphi_n)$.
- Thm. 5.10 **INCLUSION-EXCLUSION.** For any φ and ψ , $P(\varphi \vee \psi) = P(\varphi) + P(\psi) - P(\varphi \wedge \psi)$.
- Thm. 5.11 **CONJUNCTION.** For any φ and ψ : (a) if $P(\psi) \neq 0$, then $P(\varphi \wedge \psi) = P(\varphi | \psi) \times P(\psi)$; (b) if $P(\psi) = 0$, then $P(\varphi \wedge \psi) = 0$.
- Thm. 5.12 **CONJUNCTION OF INDEPENDENTS.** If φ is independent of ψ , then $P(\varphi \wedge \psi) = P(\varphi) \times P(\psi)$.
- † **PRODUCT TEST.** If $P(\psi) \neq 0$, then φ is independent of ψ iff $P(\varphi \wedge \psi) = P(\varphi) \times P(\psi)$.
- Thm. 5.13 **SYMMETRY OF INDEPENDENCE.** If $P(\varphi) \neq 0$ and $P(\psi) \neq 0$, then φ is independent of ψ iff ψ is independent of φ .
- † **CONDITIONAL COMPLEMENT.** If $P(\psi) \neq 0$, then $P(\varphi | \psi) + P(\neg\varphi | \psi) = 1$.
- Thm. 6.3 **BAYES' THEOREM.** If $P(\varphi) \neq 0$ and $P(\psi) \neq 0$, then $P(\varphi | \psi) = P(\psi | \varphi) \times P(\varphi) / P(\psi)$.
- Thm. 6.4 **TOTAL PROBABILITY.** If $0 < P(\psi) < 1$, then $P(\varphi) = P(\varphi | \psi) \times P(\psi) + P(\varphi | \neg\psi) \times P(\neg\psi)$.
- Thm. 6.5 **BAYES' THEOREM (LTP).** If $0 < P(\varphi) < 1$ and $P(\psi) \neq 0$, then

$$P(\varphi | \psi) = \frac{P(\psi | \varphi) \times P(\varphi)}{P(\psi | \varphi) \times P(\varphi) + P(\psi | \neg\varphi) \times P(\neg\varphi)}.$$

† Not proved in the text; it follows quickly from the definitions and theorems above.

Appendix: Fraction Rules

addition, same denominator: $\frac{a}{b} + \frac{c}{b} = \frac{a+c}{b}$

addition, different denominators: $\frac{a}{b} + \frac{c}{d} = \frac{a \times d}{b \times d} + \frac{c \times b}{d \times b} = \frac{a \times d + c \times b}{b \times d}$

subtraction, same denominator: $\frac{a}{b} - \frac{c}{b} = \frac{a-c}{b}$

subtraction, different denominators: $\frac{a}{b} - \frac{c}{d} = \frac{a \times d}{b \times d} - \frac{c \times b}{d \times b} = \frac{a \times d - c \times b}{b \times d}$

multiplication: $\frac{a}{b} \times \frac{c}{d} = \frac{a \times c}{b \times d}$

division: $\frac{a}{b} \div \frac{c}{d} = \frac{a/b}{c/d} = \frac{a}{b} \times \frac{d}{c} = \frac{a \times d}{b \times c}$

cancelling common factors: $\frac{a \times c}{b \times c} = \frac{a}{b}$

negatives: $-\frac{a}{b} = \frac{-a}{b}$

identity: $a = \frac{a}{1}$

percentages: $x\% = \frac{x}{100}$

Reducing fractions

A fraction is in *lowest terms* when its numerator and denominator have no common factor other than 1. To reduce a fraction to lowest terms:

1. Find a number (other than 1) that divides both the numerator and the denominator evenly. If both end in 0, try 10; otherwise try 2, 3, 5, 7, and so on.
2. Divide both by that number (this is cancelling a common factor).
3. Repeat until no such number is left.

For example, $\frac{18}{24} = \frac{9}{12} = \frac{3}{4}$ (divide by 2, then by 3). Dividing by the largest common factor, here 6, gets there in one step. Percentages reduce the same way: $80\% = \frac{80}{100} = \frac{4}{5}$.

Two shortcuts save work in probability problems:

- **Cancel before you multiply.** In $\frac{4}{9} \times \frac{3}{8}$, divide the 4 and the 8 by 4, and the 3 and the 9 by 3: $\frac{1}{3} \times \frac{1}{2} = \frac{1}{6}$. This is easier than computing $\frac{12}{72}$ and then reducing it.
- **Clear fractions inside a fraction** by multiplying its top and its bottom by the same number (a common denominator of the small fractions). In the LTP version of Bayes' theorem, multiplying top and bottom by 500 gives:

$$\frac{\frac{9}{10} \times \frac{1}{50}}{\frac{9}{10} \times \frac{1}{50} + \frac{1}{10} \times \frac{49}{50}} = \frac{\frac{9}{500}}{\frac{9}{500} + \frac{49}{500}} = \frac{9}{9 + 49} = \frac{9}{58}.$$

Index

- additivity, 54
- argument, 2
- backwards induction, 167
- best response, 153
- centipede game, 168
- combined truth-table, 38
- common knowledge, 144
- conclusion, 2
- conclusion indicators, 40
- conditional
 - indicative, 26
 - subjunctive, 27
- conditional probability, 51
- contingency, 14
- contradiction, 14
- conventional implicature, 25
- counterexample, 40
- discourse connectives, 25
- disjunction
 - exclusive, 25
 - inclusive, 25
- disjunctive normal form, 17
- dominance
 - crossing, 100
 - equivalent, 100
 - norm, 100
 - strong, 100
 - weak, 100
- dominance solvable, 150
- dominant strategy, 161
- expectation, 114
- expected utility, 117
 - conditional, 121
 - conditional norm, 121
 - norm, 117
- extensive form, 138
- fair (process), 56
- functional completeness, 18
- game, 137
- impossibility
 - logical, 32
 - nomological, 32
 - technological, 32
- independence, 55
- information set, 139
- interactive decision, 137
- interval scales, 113
- iterated deletion of strongly
 - dominated strategies, 149
- main connective, 8, 9, 17
- maximin, 156

- leximin, 103
- norm, 103
- meta-linguistic variables, 7
- minimax, 156
- mixed strategy, 155
- mutual knowledge, 144
- mutually exclusive, 50
- NAND, *see* Sheffer stroke
- Nash equilibrium, 154
- necessary proposition, 14
- non-negativity, 54
- normality, 54
- norms for credences
 - diachronic, 71
 - synchronic, 70
- parentheses-dropping conventions, 11
- Pareto optimality, 162
- partition, 94
- payoff matrix, 137
- PL
 - interpreted, 11
 - uninterpreted, 11
- positive linear transformations, 113
- possibility
 - logical, 32
 - metaphysical, 33
 - nomological, 32
 - technological, 32
- possible world, 14, 38
- preferences
 - asymmetric, 95
 - connected, 96
 - transitive, 95
 - weak order, 97
- premise, 2
- premise indicators, 40
- proposition, 3
 - atomic, 5
 - molecular, 5
- propositional logic (PL), 3
- pure strategy, 155
- ratio scales, 113
- regret, 105
 - norm, 105
- repeated game, 166
- saddle point, 156
- scope, 9
- Sheffer stroke, 18
- sound argument, 35
- strategy, 141
- strictly competitive, 145
- syntax of PL, 7
 - base clause, 7
 - closure clause, 8
 - recursive clause, 7
- syntax trees, 8
- tautology, 14
- unbiased, 56
- utility
 - ordinal, 98
 - ordinal transformation, 98
 - positive linear transformation, 113
- validity, 3, 31
 - tautological, 3, 35, 38
 - vacuous, 34
- wff, 7
- zero-sum, 145