Formal Methods in Philosophy

Lecture Notes — 2013

Dr. Anders J. Schoubye · School of Philosophy, Psychology and Language Sciences
University of Edinburgh
anders.schoubye@ed.ac.uk · http://www.schoubye.net

Contents

Preface

These notes were written for my FORMAL METHODS IN PHILOSOPHY class taught at the University of Edinburgh in the spring 2013. I owe the idea for the course to Sarah Moss who currently teaches a quite similar course at the University of Michigan. The notes will be further revised and expanded when I teach the course again in the fall of 2013.

The notes are based almost entirely on the texts listed below and they are supposed to merely function as a supplement to these texts.

- Barwise and Etchemendy (1999)
- Sider (2010)
- Weatherson (2011)

On pains of potential paradox, I am convinced that the notes contain multiple errors and typos. Consequently, I recommend taking due care when using them.

NB! This is not original research and hence should not be cited as such. Please refer to the above texts for relevant citations.

Anders J. Schoubye
16. March, 2013

Chapter 1

Summary: First Order Logic

1.1 First Order Logic (FOL)

- We start today with a recap of the syntax and semantics of **First Order Logic** (FOL). However, first we need to provide a vocabulary (or a lexicon) for our formal language $\mathfrak{L}$.

1.1.1 Primitive Vocabulary of $\mathfrak{L}$

- The language $\mathfrak{L}$ contains the following primitive expression types.
  - **Connectives**: $\neg, \wedge, \vee, \rightarrow, \leftrightarrow$
  - **Variables**: $x, y, z, \ldots$
  - **Individual Constants**: $a, b, c, \ldots$
  - **1-Place Predicates**: $F^1, F^2, F^3 \ldots F^n$
  - **2-Place Predicates**: $R^1, R^2, R^3 \ldots R^n$
  - ***n*-Place Predicates**: $G^1_n, G^2_n, G^3_n \ldots G^{\mathrm{n}}_n$
  - **Quantifiers**: $\forall, \exists$
  - **Parentheses**: (,)

We will add numerical superscripts to variables and constants when we need more than the alphabet provides.

- Variables and constants are both referred to as *terms*.

1.1.2 Syntax of $\mathfrak{L}$

- Next, we state the syntactic rules of $\mathfrak{L}$, i.e. the rules that determine whether a string of $\mathfrak{L}$ is a well formed formula (wff).

  1. If Π is an n-place predicate and $\alpha_1 \ldots \alpha_n$ are terms, then $\Pi\alpha_1 \ldots \alpha_n$ is wellformed formula (wff).
  2. If ϕ and ψ are wffs, and α is a variable, then the following are wffs:

$$\neg\phi \quad | \quad \phi \wedge \psi \quad | \quad \phi \vee \psi \quad | \quad \phi \rightarrow \psi \quad | \quad \phi \leftrightarrow \psi \quad | \quad \forall\alpha\phi \quad | \quad \exists\alpha\phi$$

3. Only strings formed on the basis of 1. and 2. are wffs.

1.1.3 Variable Binding

- A formula is *closed* if and only if all of its variables are bound. The notion of binding is defined as follows.

DEFINITION: BINDING
A variable α in a wff ϕ is bound in ϕ iff α is within an occurrence of some wff of the form $\forall\alpha\psi$ or $\exists\alpha\psi$ within ϕ. Otherwise α is free.

- For example, in the formulas below, x is bound but y is free.

$$\exists x\big(F(x)\big) \quad | \quad \exists x\big(F(y)\big) \quad | \quad \forall x\big(R(x,y)\big) \quad | \quad \forall x\big(F(x) \wedge G(y)\big)$$

- Notice that both open and closed formulas count as wffs
- Finally, notice that $\forall$ and $\exists$ are duals, so:

$$\forall\alpha\phi \leftrightarrow \neg\exists\alpha\neg\phi \quad | \quad \exists\alpha\phi \leftrightarrow \neg\forall\alpha\neg\phi$$

1.1.4 Semantics and Models for $\mathfrak{L}$

- Next, we need a semantics for $\mathfrak{L}$, i.e. an assignment of meanings to atomic expressions and a systematic method for determining the conditions under which the wffs of $\mathfrak{L}$ are true or false. This requires a model for $\mathfrak{L}$.
- A **model** $\mathfrak{M}$ is an ordered pair $\langle\mathcal{D}, \mathcal{F}\rangle$, where:
  - $\mathcal{D}$ is a non-empty set (the domain).
  - $\mathcal{F}$ is an interpretation function which satisfies the following two conditions:
    1. if α is a constant, then $\mathcal{F}(\alpha) \in \mathcal{D}$.
    2. if Π is an n-place predicate, then $\mathcal{F}(\Pi)$ is an n-place relation over $\mathcal{D}$.
- In short, the model provides an extension (i.e. meaning) of the non-logical constants, viz. individual constants and predicate constants.
- Next, we require a **recursive definition of truth** for the wffs of $\mathfrak{L}$, but since our vocabulary now includes variables and quantifiers, we need to say something about the interpretation of these expressions.

1.1.5 Variables in $\mathfrak{L}$

- A variable assignment g for a model $\mathfrak{M}$ is a function from variables in the object language $\mathfrak{L}$ to objects in the domain $\mathcal{D}$. I.e. let VAR be the set of variables, then:

$$g\colon VAR \longmapsto \mathcal{D}$$

- Here is an example of a variable assignment g:

$$g = \begin{bmatrix} x & \longrightarrow & \text{Bob} \\ y & \longrightarrow & \text{Sue} \\ z & \longrightarrow & \text{Mary} \\ \vdots & & \vdots \end{bmatrix}$$

- Hence, we should distinguish between the interpretation of constants and variables.
- Constants are interpreted relative to the interpretation function $\mathcal{F}$ in $\mathfrak{M}$ whereas variables are interpreted relative to a variable assignment g.
- Let $[\![\alpha]\!]_{\mathfrak{M},g}$ stand for the denotation of α relative to $\mathfrak{M}$ and g. So,

$$[\![\alpha]\!]^{\mathfrak{M},g} = \begin{cases} \mathcal{F}(\alpha) & \text{if } \alpha \text{ is a } \textit{constant} \\ g(\alpha) & \text{if } \alpha \text{ is a } \textit{variable} \end{cases}$$

1.1.6 Valuations and Truth-in-a-Model

- A valuation function $\mathcal{V}$ for a model $\mathfrak{M}$ and some variable assignment g is a function which assigns to each wff either 0 or 1 under the following constraints.
    - For any n-place predicate Π and any terms $\alpha_1 ... \alpha_n$, $\mathcal{V}_{\mathfrak{M},g}(\Pi\alpha_1 ... \alpha_n) = 1$ iff $\langle [\alpha_1]_{\mathfrak{M},g} ... [\alpha_n]_{\mathfrak{M},g} \rangle \in \mathcal{F}(\Pi)$
    - For any wffs ϕ, ψ, and any variable α:

$$\begin{array}{rll}
\mathcal{V}^{\mathfrak{M},g}(\neg\phi) & = 1 \text{ iff} & \mathcal{V}^{\mathfrak{M},g}(\phi) = 0 \\
\mathcal{V}^{\mathfrak{M},g}(\phi \wedge \psi) & = 1 \text{ iff} & \mathcal{V}^{\mathfrak{M},g}(\phi) = 1 \textbf{ and } \mathcal{V}^{\mathfrak{M},g}(\psi) = 1 \\
\mathcal{V}^{\mathfrak{M},g}(\phi \vee \psi) & = 1 \text{ iff} & \mathcal{V}^{\mathfrak{M},g}(\phi) = 1 \textbf{ or } \mathcal{V}^{\mathfrak{M},g}(\psi) = 1 \\
\mathcal{V}^{\mathfrak{M},g}(\phi \rightarrow \psi) & = 1 \text{ iff} & \mathcal{V}^{\mathfrak{M},g}(\phi) = 0 \textbf{ or } \mathcal{V}^{\mathfrak{M},g}(\psi) = 1 \\
\mathcal{V}^{\mathfrak{M},g}(\forall\alpha\phi) & = 1 \text{ iff} & \textbf{for every } u \in \mathcal{D}, \mathcal{V}^{\mathfrak{M},g^{[u/\alpha]}}(\phi) = 1 \\
\mathcal{V}^{\mathfrak{M},g}(\exists\alpha\phi) & = 1 \text{ iff} & \textbf{for at least one } u \in \mathcal{D}, \mathcal{V}^{\mathfrak{M},g^{[u/\alpha]}}(\phi) = 1
\end{array}$$

- We now define truth-in-a-model as follows.

DEFINITION: TRUTH-IN-A-MODEL
ϕ is *true in* a model $\mathfrak{M}$ iff $\mathcal{V}^{\mathfrak{M},g}(\phi) = 1$, for each variable assignment g for $\mathfrak{M}$.

1.1.7 Validity and Logical Consequence

- Validity is defined as truth in all models.

DEFINITION: VALIDITY
ϕ is valid in $\mathfrak{L}$ iff ϕ is true in all models $\mathfrak{M}$. $\vDash_{\mathfrak{L}} \phi$

- Logical consequence is defined in terms of truth preservation.

DEFINITION: LOGICAL CONSEQUENCE
ϕ is a logical consequence of a set of wffs Γ in $\mathfrak{L}$ iff:
For every model $\mathfrak{M}$ and every variable assignment g for $\mathfrak{M}$, if $\mathcal{V}^{\mathfrak{M},g}(\gamma) = 1$ for every $\gamma \in \Gamma$, then $\mathcal{V}^{\mathfrak{M},g}(\phi) = 1$. $\Gamma \vDash_{\mathfrak{L}} \phi$

Chapter 2

Set Theory

2.1 Naïve Set Theory

- A *set* is simply a collection of objects or individuals, i.e. a collection of chairs, a collection of numbers, or a collection of both.
- To indicate that some object x is a member of a set a, we use the symbol $\in$.

$$x \in a$$

(x is a member of a)

2.1.1 Notation

- In these notes, I follow the notational conventions of Barwise and Etchemendy (1999).
  - — variables a, b, c ... range over sets.
  - — variables x, y, z ... range over both sets and objects.
- E.g. to indicate that everything is a member of some set, we write:

$$\forall x \exists a (x \in a)$$

- To state the same using only variables ranging over objects, we could introduce a special predicate for sets, *Set*, and write

$$\forall x \exists y [Set(y) \wedge x \in y]$$

2.1.2 Basic Axioms of Naïve Set Theory

- Two basic axioms characterize naïve set theory, namely the *axiom of comprehension* and the *axiom of extensionality*.
- To characterize sets, we simply use formulas of FOL. This is easier than attempting to say something more substantial about the nature of properties.

Axiom 1: Axiom of Comprehension

- The first axiom of naïve set theory — called **the axiom of comprehension** — states exactly which sets exist.

$$\exists a \forall x[x \in a \leftrightarrow P(x)]$$

- This axiom states that there is a set a whose members are all and only those objects/individuals that satisfy the formula $P(x)$.[1]
- **NB!** This is an **axiom scheme** — when $P(x)$ is replaced by some specific wff, the result is an *instance* of the axiom scheme. So, this is in fact an infinite collection of axioms.
- As we will soon discover, this axiom scheme is inconsistent and so will have to be revised.
- Since $P(x)$ can in principle be replaced by any wff of FOL, it could contain variables other than x. As a result, the axiom of comprehension should be stated as the universal closure of all potential variables occurring in the wff.

Axiom of Comprehension

Axiom of Unrestricted Comprehension
$\forall z_1 ... \forall z_n \exists a \forall x[x \in a \leftrightarrow P(x)]$

- For example, $P(x)$ might be replaced by $(x = z_1 \vee x = z_2)$.

$$\forall z_1 ... \forall z_n \exists a \forall x[x \in a \leftrightarrow (x = z_1 \vee x = z_2)]$$

Axiom 2: Axiom of Extensionality

- The second axiom of naïve set theory is the axiom of extensionality. This principle states that the identity of a set is completely determined by its members — so if you know the members of a set b, then you know everything there is to know about the identity of that set.

Axiom of Extensionality

Axiom of Extensionality
$\forall a \forall b[\forall x(x \in a \leftrightarrow x \in b) \rightarrow a = b]$

- This axiom simply says that if two sets a and b have precisely the same members, then a and b are the same set, viz. identical.
- The identity of set depends only on what its members are — not the way it is described.
- One unique set can be characterized in distinct ways, for example the set of x's such that x is either G or not-F — or the set of x's such that if x is F then x is G.

$$\exists a \forall x\big(x \in a \leftrightarrow (\neg F(x) \vee G(x))\big) \quad | \quad \exists a \forall x\big(x \in a \leftrightarrow (F(x) \rightarrow G(x))\big)$$

[1] To avoid certain complications, we assume that the variable a cannot occur in the wff $P(x)$

Uniqueness Theorem

- From the axiom of unrestricted comprehension and the axiom of extensionality, we can prove that each wff in FOL characterizes a *unique* set, i.e. we can prove that the following is a consequence (a theorem) of comprehension and extensionality.

$$\forall z_1 ... \forall z_n \exists ! a \forall x[x \in a \leftrightarrow P(x)]$$

Read ⌜$\exists ! \alpha$⌝ as 'there exists exactly one α ...'

Proof.

— From the axiom of comprehension, it follows that there is at least one set whose members are all and only those that satisfy the wff $P(x)$.

— Let a and b be sets whose members each satisfy $P(x)$, viz.

$$\forall x[x \in a \leftrightarrow P(x)] \wedge \forall x[x \in b \leftrightarrow P(x)]$$

— It then follows

$$\forall x[x \in a \leftrightarrow x \in b]$$

— But then from the axiom of extensionality, it now follows that $a = b$.

QED

- The uniqueness theorem entails that for any wff of FOL, there is a unique set whose members are all on only those that satisfy the wff, so we can characterize these sets informally as follows.

$$\{x \mid P(x)\}$$

- This is informal notation since it is not part of our official first order language, but we can use it when convenient.
- **NB!** Everything stated using the brace notation above could also be stated using official first order language. For example, $b \in \{x \mid P(x)\}$ could be written as follows.

$$\exists a[\forall x(x \in a \leftrightarrow P(x)) \wedge b \in a]$$

SUMMING UP
Axiom of Comprehension states every first order formula determines a set.
Axiom of Extensionality states that sets with the same members are identical.

2.1.3 Empty Set, Singleton Sets, and Pairs

- Consider the wff below.

$$x \neq x$$

- Since no object satisfies this formula, the set characterized by such a formula is *the empty set*, normally denoted by $\varnothing$.

The empty set

$$\{x \mid x \neq x\} = \varnothing$$

- If only one object x satisfies $P(x)$, comprehension and extensionality together guarantee that there is a unique set that contains this object. This set is called *the singleton set containing* x.

Singleton sets

$$\{x\}$$

- Using the axioms of comprehension and extensionality, we can prove the existence of sets with exactly two members — sometimes called the **unordered pair theorem**.

Sets of pairs

- That is, we can prove the following:

$$\forall x \forall y \exists ! a \forall z (z \in a \leftrightarrow (z = x \vee z = y))$$

Proof.

— Let x and y be arbitrary objects and let $a = \{z \mid z = x \vee z = y\}$

— Since $(z = x \vee z = y)$ is a wff, existence of a is guaranteed by the axiom of comprehension.

— That a is unique follows from extensionality.

— Hence a has exactly x and y as members.

QED

- This proof strategy could of course be used to prove the existence of sets containing only one object (singleton sets), or sets with three, four, five members etc.

2.1.4 Subsets

- Given sets a and b, a is a *subset* of b, $a \subseteq b$, iff every member of a is also a member of b. For example,
  - · Let $a = \{1,2,3\}$ and let $b = \{1,2,3,4,5,6\}$.
  - · In that case, $a \subseteq b$.

Subset $\subseteq$

- The subset relation is not derived but simply defined.

DEFINITION: SUBSETS
a is a subset of b iff for all objects x, if x is a member of a, then x is a member of b.

$$a \subseteq b \leftrightarrow \forall x[x \in a \rightarrow x \in b]$$

- The following theorem is a consequence of the definition above.

$$\forall a(a \subseteq a)$$

- It can also be proved that for all sets a and b, $a = b$ iff $a \subseteq b$ and $b \subseteq a$, viz.

$$\forall a \forall b(a = b \leftrightarrow (a \subseteq b \wedge b \subseteq a))$$

Home work exercise: prove this.

2.1.5 Intersection and Union

- The **intersection** of sets a and b is the set of elements that are members of both a and b.
- We use the symbol '∩' to indicate intersections and it is defined as follows.

Intersection ∩

$$\forall a \forall b \forall z(z \in a \cap b \leftrightarrow (z \in a \wedge z \in b))$$

- The **union** of sets a and b is the set of elements that are either members of a or members of b (or both).
- We use the symbol '∪' to indicate unions and it is defined as follows.

Union ∪

$$\forall a \forall b \forall z(z \in a \cup b \leftrightarrow (z \in a \vee z \in b))$$

- Again, it can be proved from comprehension and extensionality that for any two sets a and b, there is a set which is the intersection of a and b and a set which is the union of a and b.
- Here is a proof for intersection.

Proof. (intersection)

— Let a and b be arbitrary sets. We need to prove that for such two arbitrary sets, there will always be a set c which consists of the elements that are members of both a and b.

— Consider the uniqueness theorem above — this was proved from comprehension and extensionality.

— This theorem showed that for any wff $P(x)$, there is a unique set containing all and only those elements that satisfy $P(x)$.

— So, suppose that $P(x)$ is the wff $(x \in a \wedge x \in b)$.

— In that case, it follows that $\forall a \forall b \exists! c \forall x[x \in c \leftrightarrow (x \in a \wedge x \in b)]$

QED

- The corresponding proof for *union* should be obvious.

Home work exercise: prove it for union.

SUMMING UP

$a \subseteq b$ iff every member of a is a member of b.

$x \in a \cap b$ iff $x \in a \wedge x \in b$.

$x \in a \cup b$ iff $x \in a \vee x \in b$.

2.1.6 Ordered Pairs

Ordered Pairs

- To make set theory useful for modelling structures, we need a way to represent order, i.e. we need *ordered sets*.
- Given our current axioms, sets are unordered, for example from the axiom of extensionality, it follows that:

$$\{2,1\} = \{1,2\}$$

- Henceforth, we will assume that $\langle x,y\rangle$ is a set containing two elements, namely x and y, but that this set is ordered so that x is designated as the first element and y is designated as the second.
- Ordered pairs such as $\langle x,y\rangle$ are defined as follows.

DEFINITION: ORDERED PAIRS
For any object x and y, the ordered pair $\langle x,y\rangle$ is the set $\{\{x\},\{x,y\}\}$.

$$\forall x\forall y\ \langle x,y\rangle = \{\{x\},\{x,y\}\}$$

- Since $\{\{x\}, \{x,y\}\} \neq \{\{y\},\{y,x\}\}$, this definition entails the following.

$$\langle x,y\rangle \neq \langle y,x\rangle$$

- Given this definition, it is now also to straightforward to represent ordered triples, quadruples, etc — e.g.

$$\langle x,\langle y,z\rangle\rangle \ \ldots\ \langle x,\langle y,\langle z,w\rangle\rangle\rangle \ \ldots\ \langle x, \langle\ \ldots\ \rangle\rangle$$

2.1.7 Cartesian Products

Cartesian Products

- From comprehension and extensionality, we can prove that for any two sets a and b, there is a set consisting of all the ordered pairs where the first coordinate of each pair is an element of a and the second coordinate of each pair is an element of b (we won't prove this here).
- This set is called the cartesian product of a and b — normally written as $a\times b$.
- For example, let a and b denote the sets below.

$$a = \{1,3\} \qquad\qquad b = \{2,4\}$$

- The cartesian product of a and b then equals:

$$a\times b = \{\langle 1,2\rangle, \langle 1,4\rangle, \langle 3,2\rangle, \langle 3,4\rangle\}$$

- We will say that a set c is a relation between two sets, e.g. a and b, iff it is a subset of the cartesian product of two sets, e.g. $a\times b$.

2.1.8 Relations

Relations

- Having ordered sets in our inventory, we can now represent sets containing objects that are related in certain ways.
- For example, given some domain D of objects, some predicates will express binary relations R between these objects. Such relations can be modeled by means of a set of ordered pairs, e.g.

$$\{\langle x,y\rangle \mid x \in D, y \in D, \text{and } R(x,y)\}$$

- In a first order language such as $\mathfrak{L}$, the above set is the extension of R. More specifically, a relation R is a subset of the cartesian product of two sets.

Properties of Relations

Properties of Relations

- There is an infinite number of different relations, but some relations have properties that it is useful to classify.
- For example, a relation R is said to have the property of *transitivity* if it satisfies the following condition.

$$\forall x \forall y \forall z[(R(x,y) \land R(y,z)) \to R(x,z)]$$

- I.e. for all objects x,y,z: if x is R-related to y and y is R-related to z, then x is R-related to z.
- For a more intuitive example, consider the natural language expressions ‘taller than’ and ‘father of’.
  - If Irene is taller than Angelika and Angelika is taller than Polly, then Irene is taller than Polly — so ‘taller than’ is transitive.
  - if John is the father of Stephen and Stephen is the father of Herman, then John is not the father of Herman — so ‘father of’ is not transitive.

- List of important properties of relations

DEFINITION: PROPERTIES OF RELATIONS

Transitive	$\forall x \forall y \forall z[(R(x,y) \wedge R(y,z)) \rightarrow R(x,z)]$
Intransitive	$\forall x \forall y \forall z[(R(x,y) \wedge R(y,z) \rightarrow \neg R(x,z)]$
Non-Transitive	$\exists x \neg \forall y \forall z[(R(x,y) \wedge R(y,z)) \rightarrow R(x,z)]$
Reflexive	$\forall x[R(x,x)]$
Irreflexive	$\forall x[\neg R(x,x)]$
Non-Reflexive	$\exists x \neg [R(x,x)]$
Symmetric	$\forall x \forall y[R(x,y) \rightarrow R(y,x)]$
Asymmetric	$\forall x \forall y[R(x,y) \rightarrow \neg R(y,x)]$
Antisymmetric	$\forall x \forall y[(R(x,y) \wedge R(y,x)) \rightarrow x = y]$
Non-Symmetric	$\exists x \neg \forall y[R(x,y) \rightarrow R(y,x)]$

  - The relation ‘larger than’ is *transitive*, *asymmetric*, and *irreflexive*.
  - The relation ‘father of’ is *intransitive*, *asymmetric*, and *irreflexive*.
  - The relation ‘adjoins’ is *symmetric*, *irreflexive*, and *non-transitive* (viz. does not satisfy the conditions for a transitive relation).

Inverse Relations

- Given a binary relation $R \subseteq a \times b$, we can define the inverse of R (written as R^{-1}) as follows.

$$R^{-1} = \{\langle x,y \rangle \mid \langle y,x \rangle \in R\}$$

As examples, the inverse of the relation ‘taller than’ corresponds to the meaning of the predicate ‘smaller than’. Similarly, the inverse of the predicate ‘father of’ corresponds to the meaning of the predicate ‘son or daughter of’.

Equivalence Relations and Equivalence Classes

- If a relation is reflexive, symmetric, and transitive it is known as an **equivalence relation**.
- Equivalence relations—as the name would indicate—express some kind of equivalence among the objects in the set, for example
  - ‘being identical’ (=)
  - ‘being the same size’
  - ’having the same birthday’
- Equivalence relations group objects together that are the same along one or more dimensions.

- These groupings can be modeled using what is known as **equivalence classes**.
- Given an equivalence relation $R \subseteq a{\times}b$, we can group together the objects that are deemed equivalent by R, i.e.

$$\text{Let } [x]_R \text{ be the set: } \{y \mid \langle x,y\rangle \in R\}$$

- That is, $[x]_R$ is the set of objects/individuals that are equivalent to x with respect to the (equivalence) relation R.

Functions

- A function f is a mapping from one set a to another set b, i.e. a special kind of relation $R \subseteq a{\times}b$. (what is special about it will be explicated below).
- A function f is thus a set of ordered pairs — a subset of $a{\times}b$.
    - If f is a function, viz. a subset of $a{\times}b$, then a is the *domain* of the function.
    - If f is a function, viz. a subset of $a{\times}b$, then b is the *range* of the function.
- For a relation R to be a function, it must satisfy the following condition:

Functionality: $\forall x \exists^{\leq 1} y[R(x,y)]$

Read ⌜$\exists^{\leq 1} y$⌝ as 'there is at most one y such that ...'

- So, a relation $R \subseteq a{\times}b$ is a function f if for any object in the domain of f there is a mapping to *at most* one object in the range of f.
- If the function satisfies the extra condition below, it is said to be a **total** function.

Totality: $\forall x \exists y[R(x,y)]$

total functions

- However, notice that for a relation to be a function, **it must satisfy functionality**! If the relation maps any object x to more than one output, the relation is not a function at all.
- If a function f is not total, it is a **partial** function.

partial functions

Power Sets

- For any set a, there is a unique set of all the subsets of a. This is standardly called the power set of a — written $\mathcal{P}(a)$.
- That every set has a power set can be proved from comprehension and extensionality, viz.

$$\forall a \exists b \forall x(x \in b \leftrightarrow x \subseteq a)$$

power set

Proof.

— Since $(x \subseteq a)$ is a wff, comprehension allows us to form the set:

$$b = \{x \mid x \subseteq a\}$$

— This is the set of all subsets of a.

— By extensionality, we can prove that there can be only one such set.

QED

- Consider the set below.

$$a = \{1,2,3\}$$

- The powerset of a equals the set of every subset of a, viz.

$$\mathcal{P}(a) = \{\{1\},\{2\},\{3\},\{1,2\},\{1,3\},\{2,3\},\{1,2,3\},\varnothing\}$$

- Notice that the empty set, $\varnothing$, is a member of the powerset of a.
- In general, if some set a contains n elements, the power set of a will contain 2^n elements.
- **Subsets as Members**
  Notice that it is possible for a set to have some of its elements as subsets, e.g. the sets a and b below.

$$a = \{1, \{1\}\} \qquad b = \{\text{Tom, Harry, Jack, }\{\text{Tom, Harry}\}\}$$

- However, note also that no set can have *all* of its subsets as members, i.e. it is never the case that for some set a: $\mathcal{P}a \subseteq a$.

Proof.

— We want to prove that $\mathcal{P}(a) \nsubseteq a$ — i.e. that there is a member of $\mathcal{P}(a)$ that is not a member of a.

— By the axiom of comprehension, we can construct the following set:

$$b = \{x \mid x \in a \wedge x \notin x\}$$

— This set is a subset of a as it is defined to consist of those members x of a that satisfy a certain condition, namely $x \notin x$.

— And since b is a subset of a, it follows that $b \in \mathcal{P}(a)$.

— Using proof by contradiction, we can now show that $b \notin a$.

— Suppose for proof by contradiction that $b \in a$ and now consider whether $b \in b$ or $b \notin b$.

— If $b \in b$, then b is one of the members of a that is ruled out by b, so it follows that $b \notin b$.

— If $b \notin b$, then b is an element of a that satisfies the condition $x \notin x$, hence $b \in b$.

— So, $b \in b \leftrightarrow b \notin b$ — contradiction!

— Hence, $\mathcal{P}(a) \nsubseteq a$.

QED

- This proof shows how to find, for any set a, a subset of a that is not a member of a, namely the set:

the Russell set

$$b = \{x \mid x \in a \wedge x \notin x\}$$

- This set is sometimes referred to as *the Russell set*.
- The Russell set can be computed for any set. Consider the sets below.

$$c = \{1,2,3\} \qquad d = \{1,d\}.$$

- The Russell set for c is simply $\{1,2,3\}$. The Russell set for d is simply $\{1\}$.

SUMMING UP
The powerset of a set a is the set of all its subsets: $\mathcal{P}(a) = \{b \mid b \subseteq a\}$
For any set a, the Russell set for a — the set: $\{x \mid x \in a \wedge x \notin x\}$ — is a subset of a but not a member of a.

2.2 Russell's Paradox

- We are now in a position to see that there is a serious problem with naïve set theory.
- We proved earlier that for any set a, it is not the case that $\mathcal{P}(a) \subseteq a$.
- But using the axiom of comprehension, we can show that there is a set b such that $\mathcal{P}(b) \subseteq b$.

Proof.

— From the axiom of comprehension, we can form the set containing everything, namely d below.

$$d = \{x \mid x = x\}$$

— Since everything is self-identical, d is the universal set, it contains everything.

— But if d contains everything, then every subset of d is a member of d, and hence $\mathcal{P}(d) \subseteq d$.

— But this is inconsistent with our proof that for any set a, it is not the case that $\mathcal{P}(a) \subseteq a$.

- Another way to state this paradox is to consider the set Z below.

$$Z = \{x \mid x \notin x\}$$

- One can now derive a contradiction in the following way.

Proof.

— Either $Z \in Z$ or $Z \notin Z$.

— Since Z contains everything that is not a member of itself, then if $Z \in Z$, then $Z \notin Z$.

— Since Z contains everything that is not a member of itself, then if $Z \notin Z$, then $Z \in Z$.

— So, $Z \in Z \leftrightarrow Z \notin Z$, hence contradiction.

QED

- The consequences of this result cannot be overstated. The result shows that naïve set theory is inconsistent and so must be either abandoned or significantly revised so as to avoid inconsistency.
- In the next lecture, we will look at a revised version of set theory, now standardly called *ZFC* (named after its founders Ernst Zermelo and Abraham Fraenkel), which is designed to avoid Russell's paradox.

Chapter 3

Zermelo–Fraenkel Set Theory

3.1 Cumulative Set Theory

- Naïve set theory is inconsistent, and examining Russell's paradox, it is evident that the inconsistency is due to the **Axiom of Comprehension**.
- To see this, notice that the formula below is the negation of an instance of comprehension — yet it is also a logical truth!

$$\neg\exists a \forall x (x \in a \leftrightarrow x \notin x)$$

- Notice that the truth of the above formula does not depend in anyway on the meaning of ⌜$\in$⌝ as the following is also a first order validity.

$$\neg\exists y \forall x (R(x,y) \leftrightarrow \neg R(x,x))$$

- To make sure our set theory is consistent, we need to rethink the system.

3.1.1 The Intuitive Picture

- The now standard conception of sets is the **cumulative** conception (also called the **iterative** conception) — due to Ernst Zermelo.
- This is the general idea: Start with a basic set of elements (what Zermelo called "urelements") and form whatever sets can be formed out of these. Next, using the previously formed sets and elements as building blocks, construct new sets. Repeat this process.
- For example, suppose our only basic element is the empty set.

Ur-Element	**Stage 1**	**Stage 2**	**Stage 3**	⋯
$\varnothing$	$\{\varnothing\}$	$\{\varnothing, \{\varnothing\}\}, \{\{\varnothing\}\}$	$\{\varnothing, \{\varnothing\}, \{\varnothing, \{\varnothing\}\}, \{\{\varnothing\}\}\}$ ...	

- On this conception of sets, sets come in discrete stages. Except for the “urelements”, any set which arises at some stage n must have had its members arise at some earlier stage, $n–m$ where $m \geq 1$.
- So if a set b is constructed at stage n, then the powerset of b, $\mathcal{P}(b)$, can at the earliest be constructed at $n+m$ (again where $m \geq 1$)
- Given this conception of sets, we avoid the consequence that there is a universal set. This is easily demonstrated.

  — Assume that V is the universal set.

  — Since V is a set, it must have been formed at some stage n.

  — If so, we can only form $\mathcal{P}(V)$ at the earliest at stage $n+1$.

  — So, $\mathcal{P}(V) \notin V$.

  — And since there is something, some set, which is not in V, V cannot be the universal set.

3.1.2 The Axioms of ZFC

- Modern set theory, *ZFC*, is an axiomatic theory which is designed to capture and fully characterize the sets that there are in the cumulative hierachy. The axioms of ZFC should permit us to:

  (a) prove the existence of a basic collection of sets (from the “urelements”).

  (b) form any other set in the hierarchy that is intuitively constructible from the basic collection.

- In *pure* set theory, there is only one basic element, namely the empty set, $\varnothing$, and its existence must be justified by an axiom.

Empty Set Axiom

Empty Set

$$\exists a \forall x (x \notin a)$$

- So, in pure set theory, the empty set and whatever sets can be constructed from the empty set is all there is. However, we can still use set theory to model almost any other thing, for example we should intuitively be able to model the natural numbers, i.e. as follows:

0	1	2	3	⋯
↑	↑	↑	↑	
$\varnothing$	$\{\varnothing\}$	$\{\{\varnothing\}\}$	$\{\{\{\varnothing\}\}\}$	⋯

- Or alternatively,

0	1	2	3	$\cdots$
$\uparrow$	$\uparrow$	$\uparrow$	$\uparrow$	
$\varnothing$	$\{\varnothing\}$	$\{\varnothing,\{\varnothing\}\}$	$\{\varnothing, \{\varnothing\}, \{\varnothing,\{\varnothing\}\}\}$	$\cdots$

- The point of the axioms of ZFC is to make this intuitive picture precise, viz. to characterize the basic axioms that allow us to prove the existence of each set in the hierarchy.

1. Axiom of Extensionality

$$\forall a \forall b(\forall x(x \in a \leftrightarrow x \in b) \to a = b)$$

Axiom of Extensionality

- Since the axiom of extensionality did not give rise to any inconsistencies, this seems like a plausible axiom to maintain.

2. Axiom of Separation

- From the axiom of unrestricted comprehension it followed that for any first order wff, there is a set of containing all and only those elements that satisfy this wff. That is, comprehension permitted to us to infer the existence of any set, viz.

$$\{x \mid P(x)\}$$

- However, this allows us to form sets such as the universal set which gives rise to paradoxes. We thus need a replacement axiom for comprehension and this replacement axiom needs to ensure that we cannot construct sets that are too "big".
- One way to ensure this is to assume that a set is constructible only if it is constructed out of already existing sets — this seems consistent with the general iterative picture of sets.
- So, intuitively, if we have already constructed a set a, then for any wff of FOL, we may construct a set b which—as long as b is a subset of a—is the set of elements that satisfy the wff.

$$x \in a \wedge P(x)$$

- We capture this idea using the following axiom schema.

$$\forall z_1 \ldots \forall z_n \forall a \exists b \forall x[x \in b \leftrightarrow (x \in a \wedge P(x))]$$

Axiom of Separation

- Notice that the axiom of separation does not allow us to infer the existence of any set satisfying some formula $P(x)$ — it only permits us to *separate* out the elements of some already existing set a that satisfy some formula $P(x)$.

- So if the set a has already been constructed at some stage n, then all of a's members must have been constructed prior to n. Hence, for any subset b of a satisfying $P(x)$, we can infer the existence of b at stage n.

AXIOM OF SEPARATION: VICES AND VIRTUES
From the axiom of separation, the existence of the universal set cannot be proved — in fact, it can be proved that there is no universal set.
However, from the axiom of separation alone, one cannot prove the existence of the empty set, unions, or powersets either.
So, we need more axioms to ensure the availability of these basic operations.

3. Unordered Pair Axiom

- For any two elements there is a set that has both as members.

Unordered Pair Axiom

$$\forall x \forall y \exists a \forall z (z \in a \leftrightarrow (z = x \vee z = y))$$

- If the domain is non-empty, this axiom implies the existence of a set containing containing exactly one object (a singleton set).
- If the domain has $n \geq 2$ objects, it implies the existence of a set containing exactly two objects (for any two objects).
- Using this axiom (and the assumption that the domain is non-empty), we can also prove that there is a set that has no members (viz. that there is an empty set).
  - Given that there is a *singleton* set a, then by the axiom of separation there is a subset of a which contains all and only those elements that satisfy $x \neq x$ — the empty set.

Intersection

Intersection

- It can be proved from the axiom of separation alone that for any two sets a and b, there is a set c which is the intersection of a and b.
- Consider the axiom of separation, repeated below.

$$\forall z_1 \ldots \forall z_n \forall a \exists b \forall x [x \in b \leftrightarrow (x \in a \wedge P(x))]$$

- Assume that two sets a and c have been constructed at some earlier stage n, and substitute $P(x)$ for $x \in c$ (c is universally closed).

$$\forall z_1 \ldots \forall z_n \forall a \forall c \exists b \forall x [x \in b \leftrightarrow (x \in a \wedge x \in c)]$$

4. Union Axiom

- Given any set a of sets, the union of all the members of a, $\bigcup a$, is also a set.
- This cannot be proved from the axioms in our inventory so far, so we need to stipulate this as an axiom.

$$\forall a \exists b \forall x[x \in b \leftrightarrow \exists c(c \in a \wedge x \in c)]$$

Union Axiom

- This is a generalization of the union axiom presented earlier which allows us to form the union of a set which has infinitely many sets as members.
- Using this axiom, we can show that for any two sets a and b, $a \cup b$ exists.
- Since $a \cup b$ is the set of elements in either a or b, these elements are members of some member of $\{a,b\}$. And for any two sets a and b, we can form $\{a,b\}$ using the unordered pair axiom.

Subset

- The subset relation, $\subseteq$, is defined as before.

DEFINITION: SUBSET
$\forall x[x \in a \rightarrow x \in b] \leftrightarrow a \subseteq b$

5. Power Set Axiom

- We cannot prove that every set has a powerset from our current inventory of axioms, so we need a special powerset axiom too.

$$\forall a \exists b \forall x[x \in b \leftrightarrow x \subseteq a]$$

Powerset Axiom

- This axiom states that for every set a, there is a set b such that for all x, x is a member of b if and only if x is a subset of a.
- So, every set has a powerset.

6. Axiom of Infinity

- Since we can model, say, the natural numbers on the cumulative conception of sets and since one might think that there is a set containing e.g. the natural numbers, we need an axiom that guarantees the existence of an infinite set.

$$\exists a[\varnothing \in a \wedge \forall x[x \in a \rightarrow \{x\} \in a]]$$

Axiom of Infinity

- This axiom states that there is a set containing the empty set and which contains $\{a\}$ for every a that it contains. (we are cheating a bit here since we are now using the strictly informal $\{\cdots\}$ notation in stating an axiom).

- The set guaranteed to exists by this axiom is the following.

$$\{\varnothing, \{\varnothing\}, \{\{\varnothing\}\}, \{\{\{\varnothing\}\}\}, \ldots\}$$

- This set has as many members as there are natural numbers, viz. an infinite number of elements.

7. Axiom of Choice

- Suppose a is a set whose members are all non-empty sets. Given this, there is a function f whose domain is a and which satisfies the following:

Axiom of Choice

$$\forall x[x \in a \rightarrow f(x) \in x]$$

- That is, for all sets a of non-empty sets, there is a function which picks exactly one element from each set in a.
- Such a function is called a *choice function*.
- The axiom of choice guarantees the existence of a choice function for any set a of non-empty sets.
- The axiom of choice is not a straightforward consequence of naïve set theory, and for several years it was considered controversial. Those days are however now long gone.

8. Axiom of Regularity

- Whereas the other axioms tells us something about what sets must exist, the axiom of regularity tells us something about what sets cannot exist.
- This axiom rules out the possibility of a set which contains only itself as a member, i.e.

$$S = \{S\} \qquad \text{or} \qquad \{\{\{\ldots\}\}\}$$

- It is hard to even imagine or conceptualize what such sets would be like — especially if one has the cumulative notion of a set in the background.
- Formally, the axiom is this.

Axiom of Regularity

$$\forall b[b \neq \varnothing \rightarrow \exists y(y \in b \wedge y \cap b = \varnothing)$$

- That this axiom rules out the strange sets listed above can be proved.

Proof.

- Let a be a nonempty set.
- We want to prove that $\exists x[x \in a \wedge x \cap a = \varnothing]$.
- Pick any $b \in a$ such that b occurs the earliest in the cumulation process, i.e. for any other $c \in a$, b is constructed at least as early as c.
- If we can prove that $b \cap a = \varnothing$, we are done.
- Proof by contradiction:
  - Assume that $b \cap a \neq \varnothing$.

— Let $d \in b \cap a$.

— Since $d \in a \cap b$, it follows that $d \in a$ and that $d \in b$.

— But for $d \in b$, d must have been constructed prior to b.

— On the other hand, $d \in a$, but b was chosen so that there was no element in a constructed earlier than b — contradiction.

3.1.3 Sizes of Infinite Sets

- One worry that certain people have raised, in the wake of Russell's proof that there is no universal set, is that certain axioms allow us to derive sets that are simply "too big".
- For example, the powerset axiom permits us to derive the powerset of an infinite set, and one might think that this set is simply too big to be considered a totality.
- Remember that if a set a has n members, then $\mathcal{P}(a)$ has 2^n members.
- So, what happens if a is infinite in size?

3.1.4 Cardinality and One-to-One Correspondence

- We will adopt the convention of indicating the so-called *cardinality* of a set (the number of its members) using vertical lines, i.e. $|b|$.

$$\text{if} \qquad b = \{1,2,3\} \qquad \text{then} \qquad |b| = 3$$

- As for determining whether two sets have the same size, i.e. whether the following holds:

$$|c| = |d|$$

... we assume that c and d have the same size, just in case there is an injective and surjective (also called bijective) mapping (or function) from c to d.

  - Some Properties of Functions: Injective, Surjective, and Bijective

— Functions that never map two elements of the domain c to the same element of the range d are said to be **injective** (also called one-to-one functions or functions from c into d).

— Functions where for every element in the range d there is a mapping from one (or more) elements in the domain c are said to be **surjective** (also called functions from c onto d).

— Functions that are both injective and surjective, i.e. where for every element of d there is precisely one mapping from precisely one element in c are said to be **bijective** (also called one-to-one correspondences).

— Bijective functions are special because their inverses are also functions.

- So, the sets c and d below have the same size, since there is a bijective function with c as domain and d as range.

$$c = \{\text{Bob, Sue, Jane}\}$$
$$d = \{10{,}13{,}17\}$$

Bob	Sue	Jane
↕	↕	↕
10	13	17

- This way of measuring the sizes of sets also applies to infinite sets.
- However, infinity is a strange thing and so when measuring the sizes of infinite sets, we do get some immediately counterintuitive results — e.g. it turns out that an infinite set *is the same size* as some of its proper subsets!
- Consider for example the set of natural numbers $\mathbb{N}$ as opposed to the set of even natural numbers $\mathbb{E}$.

$\mathbb{N}$	0	1	2	3	4	5	$\cdots$	n
	↕	↕	↕	↕	↕	↕		↕
$\mathbb{E}$	0	2	4	6	8	10	$\cdots$	$2n$

- Given that there is a bijective function with $\mathbb{N}$ as domain and $\mathbb{E}$ as range, it then follows that:

$$|\mathbb{N}| = |\mathbb{E}|$$

- Many people find this somewhat counterintuitive.
- Cantor also proved another important result, namely that for any set b, the following holds:

$$|\mathcal{P}(b)| > |b|$$

- From this it follows that the cardinality of the powerset of the natural numbers is greater than the cardinality of the set of natural numbers, viz.

$$|\mathcal{P}(\mathbb{N})| > |\mathbb{N}|$$

- In other words, there are sets that are larger than sets with an infinite number of elements.
- Cantor's gave a famous proof of this, sometimes called the *diagonal argument*, but we won't work through Cantor's proof here.

Chapter 4

Modal Logic

4.1 Modal Logic: Necessity and Possibility

- Modal logic is the logic of *necessity* and *possibility*.
- In modal logic, expressions such as ‘necessarily’, ‘must’, ‘ought’, ‘possibly, ‘can’, ‘might’ are treated as logical constants.
- These constants are called *modal operators* and represented by □ and ◇.

$\Box\phi$:	“It’s necessary that ϕ.” “Necessarily ϕ.” “It must be that ϕ.” …
$\Diamond\phi$:	“It’s possible that ϕ” “Possibly ϕ” “It might be that ϕ” …

- These notions of necessity and possibility are standardly explicated in terms of *possible worlds*.

A NOTE ON POSSIBLE WORLDS

A possible world w is a *complete* and *possible* scenario, i.e. it is a way that the world (viz. the entire universe) could be. There are obviously an extremely large number of possible worlds, since there are many many ways that the world could be. However, certain scenarios are not possible. For example, there are no possible worlds where both ϕ and $\neg\phi$ are true.

- Using possible worlds, we can explicate the meaning of the modal operators as follows.

"$\Box\phi$" is true iff ϕ is true in *all* possible worlds.

"$\Diamond\phi$" is true iff ϕ is true in *at least one* possible world.

4.1.1 Modals in Natural Language

- In natural language, modal words can be used to express a quite wide variety of distinct things. Consider the word 'must':

  (1) It **must** be raining. [*given what is known*]
  EPISTEMIC MODALITY

  (2) Sue **must** go to jail. [*given the laws of society*]
  DEONTIC MODALITY

  (3) When the vector sum of all forces acting on an object is zero, the velocity of the object **must** be zero. [*given the physical laws*]
  CIRCUMSTANTIAL MODALITY

  (4) Bob **must** publish a paper. [*given his desire to get a job*]
  BOULETIC MODALITY

- As should be clear, 'must' has what is often called different modal *flavors*—i.e. it can be used to make very different modal claims.
- However, when modality is analyzed in terms of possible worlds — i.e. as quantifiers over possible worlds — we can capture these different uses of modals simply by restricting the set of worlds over which the modals quantify. For example, one *could* analyze the following modal flavors as follows (although note, these are only suggestions):

  — **Epistemic modals** quantify over the set of worlds which are compatible with what is known.

  — **Deontic modals** quantify over the set of worlds which are consistent with (i.e. obeys) the laws of society.

  — **Circumstantial modals** quantify over the set of worlds that are consistent with (i.e. obey) the physical laws.

  — **Bouletic modals** quantify over the set of worlds where the subject's desires are all satisfied.

- Beware that these are only examples. It remains a point of disagreement what the correct analysis of various modals is.
- We say that modals are *alethic* if they entail truth.

4.2 Grammar of Modal Propositional Logic (MPL)

- As part of constructing a logic of necessity and possibility, we need a formal language, i.e. we need a primitive vocabulary, a syntax, and a semantics.
- The formal language we use is a simple extension of standard propositional logic with the new connectives, $\Box$ and $\Diamond$, added.

4.2.1 Primitive Vocabulary

- **Sentence Letters** P, Q, R ...
- **Connectives**: $\rightarrow, \neg, \Box$.
- **Parentheses**: (,)

Note! P, Q, R are sentence letters, not predicate constants. We have no predicate constants in propositional logic.

- We define the remaining connectives, namely $\wedge$, $\vee$, $\leftrightarrow$, and $\Diamond$, in terms of the connectives above.

4.2.2 Syntax

- Sentence letters are wellformed formulas (wffs).
- If ϕ and ψ are wffs, then $(\phi \rightarrow \psi)$, $\neg\phi$, $\Box\phi$ are also wffs.
- Nothing else is a wff.

- The remaining connectives, including $\Diamond$, and another connective, $\strictif$, are defined as follows.
  - "$\phi \wedge \psi$" $=_{\text{def}}$ "$\neg(\phi \rightarrow \neg\psi)$"
  - "$\phi \vee \psi$" $=_{\text{def}}$ "$\neg\phi \rightarrow \psi$"
  - "$\phi \leftrightarrow \psi$" $=_{\text{def}}$ "$(\phi \rightarrow \psi) \wedge (\psi \rightarrow \phi)$"
  - "$\Diamond\phi$" $=_{\text{def}}$ "$\neg\Box\neg\phi$"
  - "$\phi \strictif \psi$" $=_{\text{def}}$ $\Box(\phi \rightarrow \psi)$

- We can now translate sentence of natural language into our formal language, e.g.

(5) It's necessary that water is H_2O. $\Box H$

(6) If Jack is a bachelor, then Jack is unmarried. $\Box(B \rightarrow \neg M)$

(7) Possibly, water is not H_2O, but necessarily, it is. $\Diamond\neg H \wedge \Box H$

(8) Necessarily, water is either H_2O or it might be *XYZ*. $\Box(H \vee \Diamond X)$

4.2.3 Semantics

- There are a variety of different modal logics (logical systems) which depending on the axioms accepted will generate differents sets of valid formulas.
- For example, if either of the formulas below are assumed to be logical truths, each will place different constraints on the semantics.

$$\Box\phi \rightarrow \phi \quad | \quad \phi \rightarrow \Diamond\Box\phi \quad | \quad \Box\phi \rightarrow \Box\Box\phi$$

4.2.4 The Problem with a Truth Functional Analysis

- The first obstacle to giving a semantics for □ and ◇ is that these cannot plausibly be analyzed *truth functional* connectives.

truth functional connectives

- A connective is truth functional under the following condition: Whenever it is combined with a sentence (or sentences) to form a new (complex) sentence ϕ, the truth value of ϕ is a function of only the truth values of its component sentence(s).
- That a truth functional analysis of modal expressions is inadequate is easily demonstrated. Look at the sentences below.

(9) It's possible that Andy Clark is in Paris. $\Diamond A$

(10) It's possible that 4 is a prime number. $\Diamond P$

- What truth table should we use for ◇?

ϕ	T	F
$\Diamond\phi$	T	?

— If 'Andy Clark is in Paris' is true, then it seems true that it is possible that Andy Clark is in Paris. Hence, predicting that $\Diamond\phi$ is true whenever ϕ is true seems unproblematic.

— But what truth value should our semantics output when the embedded sentence is false? Let's consider both options.

False: We now predict that (9) is false. However that seems wrong. After all, it appears to be perfectly possible (for all we know) that Andy is in Paris.

True: We now predict that (10) is true. However that seems wrong. After all, it seems impossible to even conceive of a world in which the number 4 is prime.

- In conclusion, we need something beyond truth tables to deal with □ and ◇.

Possible World Semantics

- We turn to *possible world semantics*. In such a semantic system, the truth of a formula is always relative to a possible world.
- Except for formulas containing □ and ◇, truth values of complex sentences are determined as a function of the truth value of the component sentences within one possible world, i.e. if w is the world of evaluation:

I sometimes indicate the world of evaluation using w*.

  - "$\neg P$" is true at w iff P is false at w.
  - "$P \rightarrow Q$" is true at w iff P is false at w or Q is true at w.

- The truth values of formulas containing □ and ◇ depend on other worlds. For example,

  - "$\Box P$" is true at w iff for all accessible worlds w', P is true at w'

- We assume that a model for MPL is equipped with a binary relation $\mathcal{R}$ over the set of possible worlds. $\mathcal{R}$ thus determines exactly which worlds are accessible from the world of evaluation (e.g. the actual world).
- Thus, a possible world w' is accessible from w iff $\langle w,w' \rangle \in \mathcal{R}$ (or alternatively when $\mathcal{R}(w,w')$ holds).

Kripke Models

- The standard models for MPL are called *Kripke models*.

Named after their inventor, Saul Kripke.

- A **Kripke model** $\mathfrak{M}$ is a tuple, $\langle \mathcal{F}, \mathcal{I} \rangle$, consisting of a frame, $\mathcal{F} = \langle \mathcal{W},\mathcal{R} \rangle$ and an interpretation function $\mathcal{I}$.
  - W is a set of possible worlds.
  - $\mathcal{R}$ is a binary (accessibility) relation defined on $\mathcal{W}$, viz. $\mathcal{R} \subseteq \mathcal{W} \times \mathcal{W}$
  - $\mathcal{I}$ is a 2-place function that assigns 0 or 1 to pairs of sentence letters (P, P^1, P^2 ... P^n) and worlds (w^1, w^2 ... w^n).

    I.e. let S be the set of all sentence letters P^1, P^2 ... P^n, then:

$$\mathcal{I}\colon (S \times \mathcal{W}) \longmapsto \{0,1\}$$

- The frame $\mathcal{F}$ provides the *structure* of the model, viz. the space of possible worlds $\mathcal{W}$ and the accessibility relations $\mathcal{R} \subseteq \mathcal{W} \times \mathcal{W}$ among the worlds.
- As is usual, the interpretation function $\mathcal{I}$ assigns semantic values to all the non-logical constants (sentence letters) relative to worlds.
- Next, we need to define a valuation function $\mathcal{V}$ which permits us to determine the truth values of simple and compound, complex sentences.
- **Valuation Function**

  Where a model $\mathfrak{M} = \langle \mathcal{F},\mathcal{I} \rangle$, a valuation function $\mathcal{V}$ for $\mathfrak{M}$, $\mathcal{V}_{\mathfrak{M}}$, is defined as the 2-place function that assigns 0 or 1 to each wff relative to each $w \in \mathcal{W}$, subject to the following constraints:
  - Where α is any sentence letter and ϕ and ψ are wffs and w is any member of $\mathcal{W}$:

DEFINITIONS: VALUATIONS

$$\begin{array}{rll}
\mathcal{V}_{\mathfrak{M}}(\alpha,w) & = 1 \text{ iff} & \mathcal{I}(\alpha,w) = 1 \\
\mathcal{V}_{\mathfrak{M}}(\neg\phi, w) & = 1 \text{ iff} & \mathcal{V}_{\mathfrak{M}}(\phi, w) = 0 \\
\mathcal{V}_{\mathfrak{M}}(\phi \to \psi, w) & = 1 \text{ iff} & \mathcal{V}_{\mathfrak{M}}(\phi,w) = 0 \text{ or } \mathcal{V}_{\mathfrak{M}}(\psi,w) = 1 \\
\mathcal{V}_{\mathfrak{M}}(\Box\phi, w) & = 1 \text{ iff} & \text{For all } w' \in \mathcal{W}, \text{ if } \mathcal{R}(w, w'), \text{ then } \mathcal{V}_{\mathfrak{M}}(\phi,w') = 1
\end{array}$$

- From these, it is straightforward to prove the following:

DEFINITIONS: EXTENDED VALUATIONS

$$\begin{array}{rll}
\mathcal{V}_{\mathfrak{M}}(\phi \wedge \psi, w) & = 1 \text{ iff} & \mathcal{V}_{\mathfrak{M}}(\phi, w) = 1 \text{ and } \mathcal{V}_{\mathfrak{M}}(\psi, w) = 1 \\
\mathcal{V}_{\mathfrak{M}}(\phi \vee \psi, w) & = 1 \text{ iff} & \mathcal{V}_{\mathfrak{M}}(\phi, w) = 1 \text{ or } \mathcal{V}_{\mathfrak{M}}(\psi, w) = 1 \\
\mathcal{V}_{\mathfrak{M}}(\phi \leftrightarrow \psi, w) & = 1 \text{ iff} & \mathcal{V}_{\mathfrak{M}}(\phi, w) = \mathcal{V}_{\mathfrak{M}}(\psi, w) \\
\mathcal{V}_{\mathfrak{M}}(\Diamond\phi, w) & = 1 \text{ iff} & \text{There is a } w' \in \mathcal{W} \text{ such that } \mathcal{R}(w, w') \wedge \mathcal{V}_{\mathfrak{M}}(\phi, w') = 1 \\
\mathcal{V}_{\mathfrak{M}}(\phi \mathbin{-\!\!3} \psi, w) & = 1 \text{ iff} & \text{For all } w' \in \mathcal{W}, \text{ if } \mathcal{R}(w, w'): \\
 & & \text{either } \mathcal{V}_{\mathfrak{M}}(\phi, w') = 0 \text{ or } \mathcal{V}_{\mathfrak{M}}(\psi, w') = 1
\end{array}$$

4.3 Modal Systems

- As mentioned earlier, depending on which formulas are taken to be axioms (or logical truths), different sets of formulas will be valid.
- These differences among modal systems can be captured in terms of constraints on the accessibility relations, so we distinguish between the following different models for modal systems where each accessibility relation has the formal feature(s) specified below.

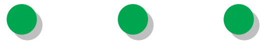
Serial Relation
That a relation $\mathcal{R}$ is *serial* means: $\forall x \exists y[\mathcal{R}(x,y)]$ — i.e. for all worlds w, there is a world w', such that w has access to w'

SYSTEM	ACCESSIBILITY RELATION
K	
D	$\mathcal{R}$ is *serial*
T	$\mathcal{R}$ is *reflexive*
B	$\mathcal{R}$ is *reflexive* and *symmetric*
S4	$\mathcal{R}$ is *reflexive* and *transitive*.
S5	$\mathcal{R}$ is *reflexive*, *symmetric*, and *transitive*

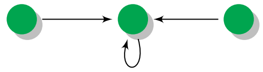
Equivalence Relation
Notice, the relation $\mathcal{R}$ required for **S5** is an *equivalence relation*, so every world has access to every world.

- Hence, any model for MPL is at least a **K**-model.

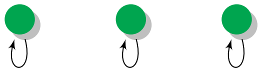
K-model (no requirements)



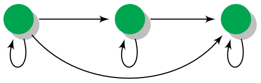
D-model (serial)



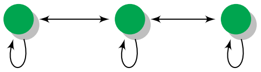
T-Model (reflexive)



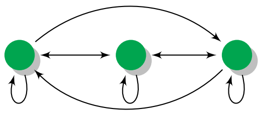
S4-model (reflexive, transitive)



B-model (reflexive, symmetric)



S5-Model (equivalence relation)



4.3.1 Validity and Consequence

- Next, we want to define validity, but since we cannot simply define validity as a formula's being true in all models (since in MPL, formulas are true or false relative to worlds within models), we need to first define a notion of validity in an MPL-model. Thus, we define validity in an MPL-model as truth in all possible worlds of the model.

DEFINITION: VALIDITY IN AN MPL-MODEL
An MPL-wff ϕ is valid in MPL-model $\mathfrak{M} = \langle \mathcal{F}, \mathcal{I} \rangle$ iff for all $w \in \mathcal{W}$, $\mathcal{V}_{\mathfrak{M}}(\phi, w) = 1$.

- We define validity and logical consequence in the standard way, viz. as truth in all models and truth preservation respectively.

DEFINITION: VALIDITY
An MPL-wff is **valid** in system Δ (where Δ is **K**, **D**, **T**, **B**, **S4**, or **S5**) iff it is valid in every Δ-model. $\vDash_\Delta \phi$

DEFINITION: LOGICAL CONSEQUENCE
MPL-wff ϕ is a **logical consequence** in system Δ of a set of MPL-wffs Γ, iff for every Δ-model $\langle \mathcal{F}, \mathcal{I} \rangle$ and every $w \in \mathcal{W}$,
if $\mathcal{V}_{\mathfrak{M}}(\gamma, w) = 1$ for each $\gamma \in \Gamma$, then $\mathcal{V}_{\mathfrak{M}}(\phi, w) = 1$ $\Gamma \vDash_\Delta \phi$

- Notice that when we use the standard notation for validity and semantic consequence ($\vDash$), it is imperative that we indicate for which system the validity holds, i.e.

  '$\vDash_{S4} \Box\phi \rightarrow \Box\Box\phi$' means that the formula '$\Box\phi \rightarrow \Box\Box\phi$' is valid in **S4**

A FURTHER NOTE ON POSSIBLE WORLDS
It's important to realize that the notion of *possible world* is completely dispensable. $\mathcal{W}$ is just a non-empty set of something — it does not have to be possible worlds. It could be numbers, people, apples, bananas etc. Similarly $\mathcal{R}$ is just a binary relation on $\mathcal{W}$ and $\mathcal{I}$ is just any old function mapping pairs of sentences and members of $\mathcal{W}$ to 0 or 1. Hence we are not obviously committed to any radical metaphysical claims.

4.4 Establishing Validities

- We are now in a position to prove that various validities hold across different modal systems. For example, we can prove the following:

  - $\vDash_K \Box(\phi \vee \neg\phi)$ (i.e. this formula is **K**-valid)
  - $\vDash_T \Box\phi \rightarrow \phi$ (i.e. this formula is **T**-valid)
  - $\vDash_{S4} \Box\phi \rightarrow \Box\Box\phi$ (i.e. this formula is **S4**-valid)

Home work exercise: Give validity proofs for the following formulas.
$\vDash_T \neg\phi \rightarrow \neg\Box\phi$
$\vDash_{S4} \neg\Box\Box\phi \rightarrow \neg\neg\Diamond\neg\phi$

- Some informal proofs:

  - **Proof**: $\vDash_K \Box(\phi \vee \neg\phi)$

(1) Suppose for reductio that $\mathcal{V}_{\mathfrak{M}}(\Box(\phi \vee \neg\phi), w) = 0$

(2) It follows that there is a world w' such that $\mathcal{R}(w,w')$ where $\mathcal{V}_{\mathfrak{M}}(\phi \vee \neg\phi, w') = 0$

(3) Given the truth table for '$\vee$', $\mathcal{V}_{\mathfrak{M}}(\phi \vee \neg\phi, w') = 0$ iff $\mathcal{V}_{\mathfrak{M}}(\phi, w') = 0$ and $\mathcal{V}_{\mathfrak{M}}(\neg\phi, w') = 0$

(4) From this is it follows that $\mathcal{V}_{\mathfrak{M}}(\phi, w') = 0$ and $\mathcal{V}_{\mathfrak{M}}(\phi, w') = 1$.

(*contradiction*)

  - **Proof**: $\vDash_{S4} \Box\phi \rightarrow \Box\Box\phi$.

(1) Suppose for reductio that $\mathcal{V}_{\mathfrak{M}}(\Box\phi \rightarrow \Box\Box\phi, w) = 0$

(2) So, $\mathcal{V}_{\mathfrak{M}}(\Box\phi, w) = 1$

(3) And, $\mathcal{V}_{\mathfrak{M}}(\Box\Box\phi, w) = 0$

(3) Since $\mathcal{V}_{\mathfrak{M}}(\Box\Box\phi, w) = 0$, then for some possible world w', such that $\mathcal{R}(w,w')$, $\mathcal{V}_{\mathfrak{M}}(\Box\phi, w') = 0$.

(5) Since $\mathcal{V}_{\mathfrak{M}}(\Box\phi, w') = 0$, then for some possible world w'' such that $\mathcal{R}(w',w'')$, $\mathcal{V}_{\mathfrak{M}}(\phi, w'') = 0$.

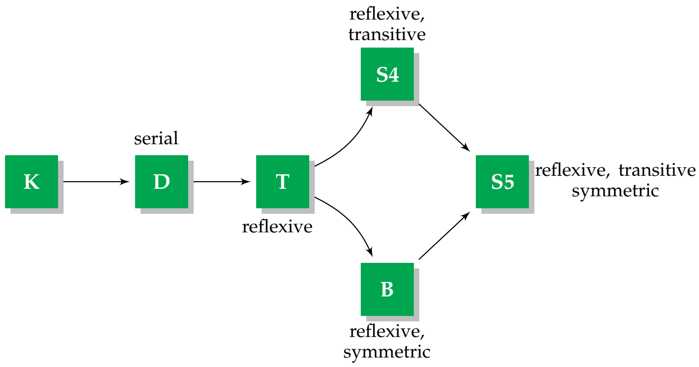
(6) Given (2), for every world w' such that $\mathcal{R}(w,w')$, $\mathcal{V}_{\mathfrak{M}}(\phi, w') = 1$

(7) **S4** is transitive, so if $\mathcal{R}(w,w')$ and $\mathcal{R}(w',w'')$, then $\mathcal{R}(w,w'')$

(8) So, $\mathcal{V}_{\mathfrak{M}}(\phi,w'') = 1$.

(9) Hence, $\mathcal{V}_{\mathfrak{M}}(\phi,w'') = 1$ and $\mathcal{V}_{\mathfrak{M}}(\phi,w'') = 0$ *(contradiction)*

- The "strength" of the modal systems mentioned above follow the ordering below.



- This means that if a wff ϕ is **K**-valid, viz. $\vDash_{\mathbf{K}} \phi$, then it follows that ϕ is valid in **D**, **T**, **B**, **S4**, and **S5** as well.
- Similarly, if a formula is **T**-valid, it follows that it is also valid in **B**, **S4**, and **S5**.
- The exceptions here are **B** and **S4**. If a formula is **S4**-valid, it follows that it is **S5**-valid. However, it does not follow that it is valid in **B**. The same point holds for **B**-validities and **S4**.
- Modal systems are defined in terms of constraints on their accessibility relation and depending on these constraints, different formulas will be valid.
- Below is a list of the modal systems that are considered here and a list of axioms that are validated by the corresponding system. The constraint that the axiom imposes on the system is indicated on the right.

MODAL SYSTEM	AXIOM	RELATIONAL PROPERTY
K–axiom	$\vDash_{\mathbf{K}} \Box(\phi \rightarrow \psi) \rightarrow (\Box\phi \rightarrow \Box\psi)$	
D–axiom	$\vDash_{\mathbf{D}} \Box\phi \rightarrow \Diamond\phi$	*serial*
T–axiom	$\vDash_{\mathbf{T}} \Box\phi \rightarrow \phi$	*reflexive*
B–axiom	$\vDash_{\mathbf{B}} \phi \rightarrow \Box\Diamond\phi$	*symmetric*
S4–axiom	$\vDash_{\mathbf{S4}} \Box\phi \rightarrow \Box\Box\phi$	*transitive*
S5–axiom	$\vDash_{\mathbf{S5}} \Diamond\phi \rightarrow \Box\Diamond\phi$	*symmetric, transitive*

4.5 Invalidity and Countermodels

- To demonstrate that a formula is invalid in some system Δ, one constructs a *countermodel* — viz. a model containing a world where the formula is false.

- To give a countermodel, one option is to write a full explication of a model. For example, suppose we wanted to demonstrate that the formula below is not **K**-valid.

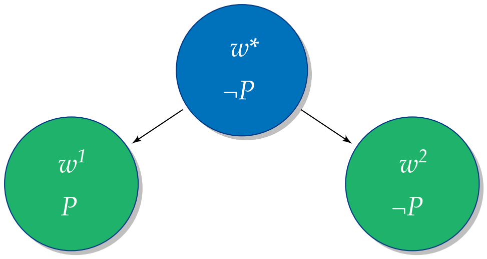
$$\nvDash_{\mathbf{K}} \Diamond P \rightarrow \Box P$$

- We construct a countermodel as follows. Let $\mathfrak{M}^1 = \langle \mathcal{F}, \mathcal{I} \rangle$ where $\mathcal{F} = \langle \mathcal{W}, \mathcal{R} \rangle$:

$\mathfrak{M}^1$	
$\mathcal{W}$:	w^*, w^1, w^2
$\mathcal{R}$:	$\mathcal{R}(w^*, w^1)$, $\mathcal{R}(w^*, w^2)$
$\mathcal{I}$:	$\langle\langle P, w^1 \rangle, 1\rangle$ (all other pairs map to 0).

- Above $\mathcal{I}$ indicates the worlds at which sentence letters and world pairs are mapped to 1. Assume that for any other pair, it is mapped to 0.

- We can illustrate the model above graphically as follows.

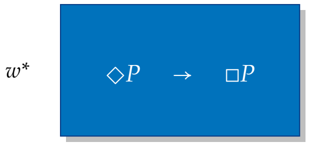


- It is easy to recognize that $\mathfrak{M}^1$ is a countermodel to the formula above.
  - To identify a countermodel, we need to find a world w where $\mathcal{V}_{\mathfrak{M}}(\Diamond P, w) = 1$ (antecedent) and $\mathcal{V}_{\mathfrak{M}}(\Box P, w) = 0$ (consequent).
  - Since $\mathcal{R}(w^*, w^1)$ and $\langle\langle P, w^1\rangle, 1\rangle \in \mathcal{I}$, it follows that $\mathcal{V}_{\mathfrak{M}}(\Diamond P, w^*) = 1$
  - But since $\mathcal{R}(w^*, w^2)$ and $\langle\langle P, w^2\rangle, 0\rangle \in \mathcal{I}$, it follows that $\mathcal{V}_{\mathfrak{M}}(\Box P, w^*) = 0$
  - Hence, $\mathcal{V}_{\mathfrak{M}}(\Diamond P \rightarrow \Box P, w^*) = 0$

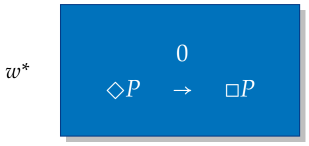
4.5.1 Graphical Procedure for Demonstrating Invalidity

- Sider uses the following graphical procedure for demonstrating invalidities.

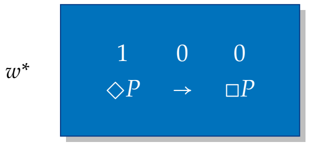
1. Enter formula in box (world).



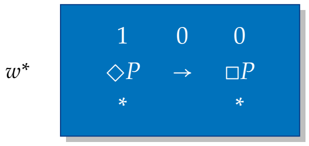
2. Enter invalidity to be established.



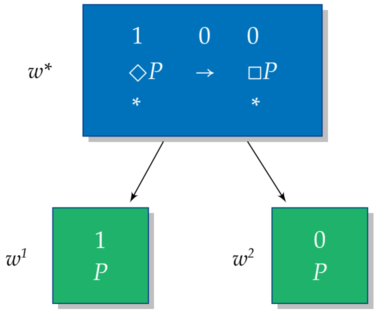
3. Enter "forced" truth values.



4. Enter asterisks.



5. Discarge bottom asterisks.



- From this, the following countermodel can now be determined:

$$\mathcal{W} = \{w^*, w^1, w^2\} \qquad \mathcal{R} = \{\langle w^*, w^1\rangle, \langle w^*, w^2\rangle\} \qquad \mathcal{I} = \{\langle\langle P, w^1\rangle, 1\rangle \ldots\}$$

- Alternative to the cumbersome notation for $\mathcal{I}$, one could also just write the following:

$$\begin{array}{lcl} w^* & : & \neg P \\ w^1 & : & P \\ w^2 & : & \neg P \end{array}$$

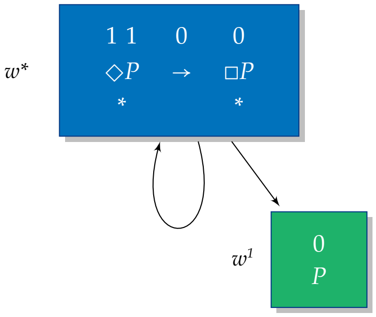
- The model above can however be simplified a bit — we do not three worlds to give a countermodel.
- In particular, if we simply let w^* access itself and assume that P is true at w^*, we get an equally suitable countermodel.

OFFICIAL MODEL

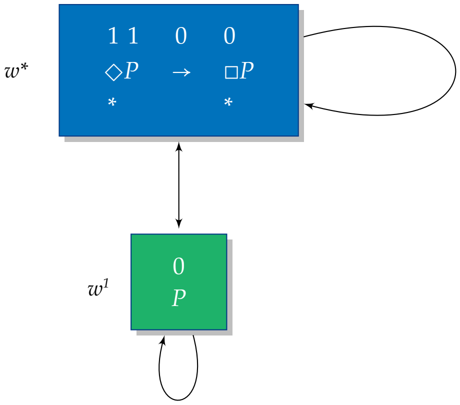
$\mathcal{W}$: $\{w^*, w^1\}$

$\mathcal{R}$: $\{\langle w^*,w^*\rangle, \langle w^*,w^1\rangle$

$\mathcal{I}$: $\{\langle\langle P,w^*\rangle, 1\rangle$ ... $\}$



- We could now proceed to show that this formula is also invalid in the more powerful systems **D**, **T**, **B**, **S4**, and **S5**.
- However, in this case it will be easier to just show that the formula is invalid in **S5**. From this it follows that it is invalid in all the other systems.

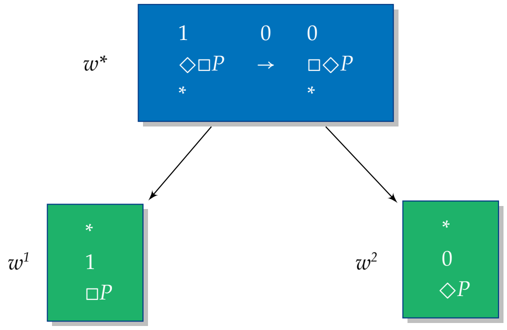


- This is an **S5**-model since for every world w (w^* and w^1), each world has access to every other world.
- But this is also a countermodel to the formula '$\Diamond\phi \rightarrow \Box\phi$'.
- Hence, this formula is invalid in every modal system (or at least in the systems that we are considering here).

- Let's consider a slightly harder case.

$$\nvDash_{\text{K}} \Diamond\Box\phi \rightarrow \Box\Diamond\phi$$

- After a few steps, we reach the representation below.

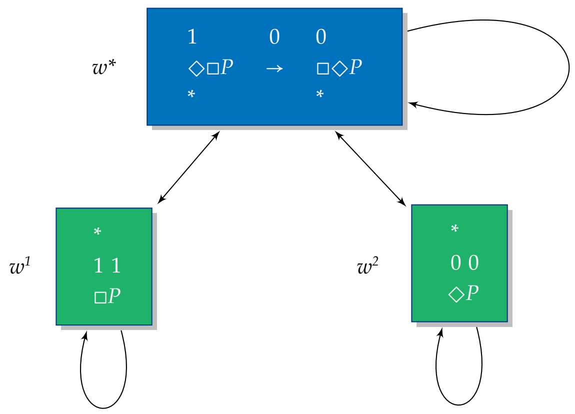


- We now need to evaluate a true $\Box\phi$-claim and a false $\Diamond\phi$-claim.
  - To determine whether $\Box\phi$ is true, we must establish whether P is true at *every* accessible world.
  - To determine whether $\Diamond\phi$ is false, we must establish whether P is false at *every* possible world.
- These are *universal* modal commitments (as opposed to the *existential* commitments we faced in the previous example) — we indicate this by placing the asterisks at the top rather than the bottom.
- As it happens, discharging the top asterisks in a **K**-model is easy. Remember:
  - · $\mathcal{I}(\Box\phi, w^1) = 1$ iff for all w', if $\mathcal{R}(w^1, w')$, then $\mathcal{I}(\phi, w') = 1$

    — But to make this true, we can simply leave the model as is since if w^1 has access to no worlds, this universal claim is vacuously true.
  - · $\mathcal{I}(\Diamond\phi, w^1) = 1$ iff there is a w' such that $\mathcal{R}(w^1, w')$ and $\mathcal{I}(\phi, w') = 1$

    — But to make this true, we can simply leave the model as is since if w^1 has access to no worlds, this existential is false.
- Hence, we have shown that there is a **K**-model in which the formula '$\Diamond\Box\phi \rightarrow \Box\Diamond\phi$' is false. Hence, the formula is not **K**-valid.

- Let's try our luck in a **B**-model.



- We now need to discharge the top asterisks, hence we need to assign truth values to P at w^*, w^1 and w^2.
- However, this is not possible without generating a contradiction.
  - Since '$\Box\phi$' is true at w^1 and $\mathcal{R}(w^1, w^*)$, P must be true in w^*.
  - But since '$\Diamond\phi$' is false at w^2 and $\mathcal{R}(w^2, w^*)$, P must be false in w^*.
- So, we're stuck.

A NOTE ON COUNTERMODELS

It's important to realize that failing to construct a countermodel in some system Δ is not sufficient for proving that the formula is valid in Δ. It only suggests that it *might* be valid.

Home work exercise: Do a validity proof of $\vDash_B \Diamond\Box P \rightarrow \Box\Diamond P$

- We can *prove* that it is impossible to give a countermodel in **B**-system by showing that the formula is true at every world in a **B**-model.

4.6 Axiomatic Proof Theory

- We now move to the proof theoretic conception of logical consequence as opposed to the semantic conception considered in the last lecture.
- On this conception, the logical consequences of a set Γ are those statements that can be proved if one takes the members of Γ as premises. A logical truth is a statement that can be proved from the empty set (i.e. from no premises).
- A proof is a step-wise reasoning process which proceeds according to clearly defined mechanical rules.
- On the axiomatic approach to proof theory (sometimes called the Hilbert approach), reasoning with assumptions is not allowed. As a result, one cannot do conditional proofs or prove formulas using reductio ad absurdum.
- Instead, an axiomatic proof is a list of formulas where each line in the proof must either be an axiom or inferred from earlier lines using an acceptable inference rule.
- Axioms should generally represent *indisputable* logical truths and from these logical truths, further logical truths (theorems) can then be proved.
- Let's consider an axiomatic proof system for modal system **K**.

4.6.1 System K

- **Inference Rules**

1. Modus Ponens

$$\frac{\phi \quad \phi \rightarrow \psi}{\psi}\,\text{MP}$$

2. Necessitation

$$\frac{\phi}{\Box\phi}\,\text{NEC}$$

AXIOMS
All instances of (with MPL-wffs) of PL1-PL3 and the **K**-axiom

PL1	$\phi \rightarrow (\psi \rightarrow \phi)$	
PL2	$(\phi \rightarrow (\psi \rightarrow \chi)) \rightarrow ((\phi \rightarrow \psi) \rightarrow (\phi \rightarrow \chi))$	
PL3	$(\neg\psi \rightarrow \neg\phi) \rightarrow ((\neg\psi \rightarrow \phi) \rightarrow \psi)$	
K	$\Box(\phi \rightarrow \psi) \rightarrow (\Box\phi \rightarrow \Box\psi)$	(distribution axiom)

TAUTOLOGIES

Useful tautologies that we will rely on in our axiomatic proofs.

DN	$\phi \leftrightarrow \neg\neg\phi$	(double negation)
CP	$(\phi \rightarrow \psi) \rightarrow (\neg\psi \rightarrow \neg\phi)$	(contraposition)
SL	$((\phi \rightarrow \psi) \wedge (\psi \rightarrow \chi)) \rightarrow (\phi \rightarrow \chi)$	(syllogism)
IE	$(\phi \rightarrow (\psi \rightarrow \chi)) \leftrightarrow ((\phi \wedge \psi) \rightarrow \chi)$	(import/export)
PM	$(\phi \rightarrow (\psi \rightarrow \chi)) \leftrightarrow (\psi \rightarrow (\phi \rightarrow \chi))$	(permutation)
CN	$((\phi \rightarrow \psi) \wedge (\phi \rightarrow \chi)) \leftrightarrow (\phi \rightarrow (\psi \wedge \chi))$	(composition)
DL	$((\phi \rightarrow \chi) \wedge (\psi \rightarrow \chi)) \leftrightarrow ((\phi \vee \psi) \rightarrow \chi)$	(dilemma)
BI	$((\phi \rightarrow \psi) \wedge (\psi \rightarrow \phi)) \leftrightarrow (\phi \leftrightarrow \psi)$	(biconditional)
DI	$(\neg\phi \rightarrow \psi) \leftrightarrow (\phi \vee \psi)$	(disjunction)
NC	$(\phi \rightarrow \neg\psi) \leftrightarrow \neg(\phi \wedge \psi)$	(negated conjunction)

- Notice that if reasoning from premises is acceptable, necessitation gets us into trouble as e.g. the following could be proved.

$$\{P\} \vdash_{\mathbf{K}} \Box P$$

- This is obviously bad as it is easy to construct a countermodel. Hence,

$$\{P\} \nvDash_{\mathbf{K}} \Box P$$

- This means that if reasoning from premises was acceptable, our system would be unsound, i.e. it would not be true that whenever $\Gamma \vdash_{\mathbf{K}} \phi$, then $\Gamma \vDash_{\mathbf{K}} \phi$.

- However, if P above is a *logical truth*, this is obviously unproblematic. Thus, when the rule of necessitation is applied, it must always be applied to an axiom or a theorem (a formula proved from the axioms and the acceptable inference rules) — not a premise.

- Also, even if there is a proof of $\Box P$ from P in **K**, it is not the case that $\vdash_{\mathbf{K}} P \rightarrow \Box P$. In effect, this means that the deduction theorem fails for all of the axiomatic systems that we are going to consider.

- The deduction theorem states that where Γ is a set of wffs and ϕ is a closed formula:

$$\Gamma \cup \{\phi\} \vdash \psi \quad \Rightarrow \quad \Gamma \vdash \phi \rightarrow \psi$$

- Since the deduction theorem is needed to do conditional proofs in axiomatic systems, we simply cannot do conditional proofs, i.e. proofs that show that a conditional is a theorem by assuming the antecedent and proving its consequent on that basis.

- Our solution is to just stay away from proofs with premises.

Example Proofs

- Let's consider some examples of axiomatic proofs.

Proof. $\Box((P \to Q) \to (P \to P))$

1.	$P \to (Q \to P)$	**PL1**
2.	$P \to (Q \to P)) \to ((P \to Q) \to (P \to P))$	**PL2**
3.	$(P \to Q) \to (P \to P)$	1, 2, **MP**
4.	$\Box((P \to Q) \to (P \to P))$	3, **NEC**

- To make things easier, we adopt the convention that whenever ψ is a tautological MPL-consequence of a set of wffs ϕ_1 ... ϕ_n, then we may simply write ψ while annotating the line numbers of ϕ_1 ... ϕ_n and **PL**.
- I.e. suppose that ψ is a tautological MPL-consequence of ϕ_1 ... ϕ_n. If so, then the following is legitimate:

1.	ϕ_1	
2.	ϕ_2	
⋮		
13.	ϕ_3	
⋮		
16.	ψ	1, 2, 13, **PL**

- A proof with **K**-axiom at work:

Proof. $\Box(P \wedge Q) \to (\Box P \wedge \Box Q)$

1.	$(P \wedge Q) \to P$	**PL**
2.	$\Box[(P \wedge Q) \to P]$	1, **NEC**
3.	$\Box[(P \wedge Q) \to P] \to [\Box(P \wedge Q) \to \Box P]$	**K**
4.	$\Box(P \wedge Q) \to \Box P$	2, 3, **MP**
5.	$(P \wedge Q) \to Q$	**PL**
6.	$\Box[(P \wedge Q) \to Q]$	5, **NEC**
7.	$\Box[(P \wedge Q) \to Q] \to [\Box(P \wedge Q) \to \Box Q]$	**K**
8.	$\Box(P \wedge Q) \to Q$	7, 8, **MP**
9.	$\Box(P \wedge Q) \to (\Box P \wedge \Box Q)$	4, 8, **PL** (composition)

- Let's attempt to prove a formula that has a necessity modal distributed over a disjunction.

Proof. $(\Box P \vee \Box Q) \to \Box(P \vee Q)$

1.	$P \to (P \vee Q)$	**PL**
2.	$\Box(P \to (P \vee Q))$	1, **NEC**
3.	$\Box(P \to (P \vee Q)) \to (\Box P \to \Box(P \vee Q))$	2, **K**
4.	$\Box P \to \Box(P \vee Q)$	2, 3, **MP**
5.	$Q \to (P \vee Q)$	**PL**
6.	$\Box(Q \to (P \vee Q))$	4, **NEC**
7.	$\Box(Q \to (P \vee Q)) \to (\Box Q \to \Box(P \vee Q))$	6, **K**
8.	$\Box Q \to \Box(P \vee Q)$	6, 7, **MP**
9.	$(\Box P \vee \Box Q) \to \Box(P \vee Q)$	4, 8, **PL** (dilemma)

- Nested necessities

Proof. $\Box\Box(P \wedge Q) \to \Box\Box P$

1.	$(P \wedge Q) \to P$	**PL**
2.	$\Box((P \wedge Q) \to P)$	1, **NEC**
3.	$\Box((P \wedge Q) \to P) \to (\Box(P \wedge Q) \to \Box P)$	**K**
4.	$\Box(P \wedge Q) \to \Box P$	2, 3, **MP**
5.	$\Box(\Box(P \wedge Q) \to \Box P)$	4, **NEC**
6.	$\Box(\Box(P \wedge Q) \to \Box P) \to (\Box\Box(P \wedge Q) \to \Box\Box P)$	**K**
7.	$\Box\Box(P \wedge Q) \to \Box\Box P$	5, 6, **MP**

- The **K**-axiom has an analogue of distributed possibility operators, namely:

K$\Diamond$	$\Box(\phi \to \psi) \to (\Diamond\phi \to \Diamond\psi)$	($\Diamond$-distribution)

- Remember '$\Diamond$' is an abbreviation for '$\neg\Box\neg$', so '$\Diamond\phi \to \Diamond\psi$' is an abbreviation for '$\neg\Box\neg\phi \to \neg\Box\neg\psi$'.

- Let's prove **K**$\Diamond$.

Proof. $\Box(\phi \to \psi) \to (\neg\Box\neg\phi \to \neg\Box\neg\psi)$

1.	$(\phi \to \psi) \to (\neg\psi \to \neg\phi)$	**PL** (**CP**)
2.	$\Box((\phi \to \psi) \to (\neg\psi \to \neg\phi))$	1, **NEC**
3.	$\Box((\phi \to \psi) \to (\neg\psi \to \neg\phi)) \to (\Box(\phi \to \psi) \to \Box(\neg\psi \to \neg\phi))$	**K**
4.	$\Box(\phi \to \psi) \to \Box(\neg\psi \to \neg\phi)$	2, 3, **MP**
5.	$\Box(\neg\psi \to \neg\phi) \to (\Box\neg\psi \to \Box\neg\phi)$	**K**
6.	$\Box(\phi \to \psi) \to (\Box\neg\psi \to \Box\neg\phi)$	4, 5, **PL** (**SL**)
7.	$\Box(\phi \to \psi) \to (\neg\Box\neg\phi \to \neg\Box\neg\psi)$	6, **PL** (contraposition)
	$\Box(\phi \to \psi) \to (\Diamond\phi \to \Diamond\psi)$	

- We can also prove the following modal negation schemas.

$\vdash_K \neg\Box\phi \to \Diamond\neg\phi$	(**MN**)	$\vdash_K \Diamond\neg\phi \to \neg\Box\phi$	(**MN**)
$\vdash_K \neg\Diamond\phi \to \Box\neg\phi$	(**MN**)	$\vdash_K \Box\neg\phi \to \neg\Diamond\phi$	(**MN**)

- Let's prove one of them.

Sider proves another one in the book **Homework Exercise**: Prove the last two (it's *extremely* easy).

Proof. $\Diamond\neg\phi \to \neg\Box\phi$

1.	$\phi \to \neg\neg\phi$	**PL** (**DN**)
2.	$\Box(\phi \to \neg\neg\phi)$	1, **NEC**
3.	$\Box(\phi \to \neg\neg\phi) \to (\Box\phi \to \Box\neg\neg\phi)$	**K**
4.	$\Box\phi \to \Box\neg\neg\phi$	2, 3, **MP**
5.	$\neg\Box\neg\neg\phi \to \neg\Box\phi$	4, **PL** (contraposition)
	$\Diamond\neg\phi \to \neg\Box\phi$	

- Moving a negation through a string of boxes and diamonds.

Proof. $\Box\Diamond\Box\neg P \to \neg\Diamond\Box\Diamond P$

1.	$\Box\neg P \to \neg\Diamond P$	**MN**
2.	$\Box(\Box\neg P \to \neg\Diamond P)$	1, **NEC**
3.	$\Box(\Box\neg P \to \neg\Diamond P) \to (\Diamond\Box\neg P \to \Diamond\neg\Diamond P)$	**K**$\Diamond$
4.	$\Diamond\Box\neg P \to \Diamond\neg\Diamond P$	2, 3, **MP**
5.	$\Diamond\neg\Diamond P \to \neg\Box\Diamond P$	**MN**
6.	$\Diamond\Box\neg P \to \neg\Box\Diamond P$	4, 5, **PL** (syllogism)
7.	$\Box(\Diamond\Box\neg P \to \neg\Box\Diamond P)$	6, **NEC**
8.	$\Box(\Diamond\Box\neg P \to \neg\Box\Diamond P) \to (\Box\Diamond\Box\neg P \to \Box\neg\Box\Diamond P)$	**K**
9.	$\Box\Diamond\Box\neg P \to \Box\neg\Box\Diamond P$	7, 8, **MP**
10.	$\Box\neg\Box\Diamond P \to \neg\Diamond\Box\Diamond P$	**MN**
11.	$\Box\Diamond\Box\neg P \to \neg\Diamond\Box\Diamond P$	9, 10, **PL** (syllogism)

4.6.2 System D

- Moving to system **D**, we add the following axiom.

D	$\Box\phi \to \Diamond\phi$

- System **D** is not much stronger than **K**, i.e. it is still not possible to prove that what is necessary is true.
- However, with the **D**-axiom, it is possible to prove e.g. that tautologies are not only necessary, but possible too.

Proof. $\Diamond(P \lor \neg P)$

1.	$P \lor \neg P$	**PL**
2.	$\Box(P \lor \neg P)$	1, **NEC**
3.	$\Box(P \lor \neg P) \to \Diamond(P \lor \neg P)$	**D**
4.	$\Diamond(P \lor \neg P)$	2, 3, **MP**

4.6.3 System T

- In system **T**, the **D**-axiom is dropped and the **T**-axiom is added instead.
- Remember, system **T** includes the **K**-axiom, **PL1**–**PL3** and all the tautologies (theorems) listed earlier.

T	$\Box\phi \to \phi$	(Factivity)

- The **T**-axiom not only gives us the quite intuitively plausible principle that what is necessary is true, but also that what is true is possible.

Note, this cannot be proved in either **K** or **D**.

Proof. $\phi \to \Diamond\phi$

1.	$\Box\neg\phi \to \neg\phi$	**T**
2.	$\neg\neg\phi \to \neg\Box\neg\phi$	1, **PL** (contraposition)
3.	$(\neg\neg\phi \to \neg\Box\neg\phi) \to (\phi \to \neg\Box\neg\phi)$	**PL** (double negation*)
4.	$\phi \to \neg\Box\neg\phi$	2, 3, **MP**

$\phi \to \Diamond\phi$

- We can refer to this theorem as **T**$\Diamond$.

T$\Diamond$	$\phi \to \Diamond\phi$

4.6.4 System B

- In system **B**, we keep axioms **K**, **T**, **PL1**–**PL3** and the various tautologies (theorems) listed above, and in addition add the following axiom.

B	$\Diamond\Box\phi \to \phi$

- With the **B**-axiom in our system, we can prove the following theorem.

B$\Diamond$	$\phi \to \Box\Diamond\phi$

Proof. $\phi \to \Box\Diamond\phi$

1.	$\Diamond\Box\neg\phi \to \neg\phi$	**B**
2.	$\neg\neg\phi \to \neg\Diamond\Box\neg\phi$	1, **PL** (contraposition)
3.	$\phi \to \neg\Diamond\Box\neg\phi$	2, **PL** (double negation)
4.	$\phi \to \Box\Diamond\phi$	**PL** (**MN**)

4.6.5 System S4

- Moving on to system **S4**, we keep axioms **K**, **T**, **PL1**–**PL3** and the various tautologies above, we add the **S4**-axiom, but no longer have axiom **B**.

S4	$\Box\phi \to \Box\Box\phi$	(Positive Introspection)

- With the **S4**-axiom we can now prove the following.

S4$\Diamond$	$\Diamond\Diamond\phi \to \Diamond\phi$

Proof. $\Diamond\Diamond\phi \to \Diamond\phi$

1.	$\Box\neg\phi \to \Box\Box\neg\phi$	**S4**
2.	$\Box\neg\phi \to \neg\Diamond\phi$	**PL** (**MN**)
3.	$\Box(\Box\neg\phi \to \neg\Diamond\phi)$	1, **NEC**
4.	$\Box(\Box\neg\phi \to \neg\Diamond\phi) \to (\Box\Box\neg\phi \to \Box\neg\Diamond\phi)$	**K**
5.	$\Box\Box\neg\phi \to \Box\neg\Diamond\phi$	3, 4, **MP**
6.	$\Box\neg\phi \to \Box\neg\Diamond\phi$	1, 5, **PL** (syllogism)
7.	$\Diamond\Diamond\phi \to \Diamond\phi$	6, **PL** (contraposition)

4.6.6 System S5

- System **S5** is the strongest of our modal systems. We get **S5** by adding the following axiom to system **T**.

S5	$\Diamond\Box\phi \to \Box\phi$	(Negative Introspection)

- In **S5**, we have neither the **S4**-axiom nor the **B**-axiom included as axioms, however they follow straightforwardly from the **S5**-axiom.
- Let's prove both.

Proof of B. $\Diamond\Box\phi \to \phi$

1.	$\Diamond\Box\phi \to \Box\phi$	**S5**
2.	$\Box\phi \to \phi$	**T**
3.	$\Diamond\Box\phi \to \phi$	1, 2, **PL** (syllogism)

Proof of S4. $\Box\phi \to \Box\Box\phi$

1.	$\Box\phi \to \Box\Diamond\Box\phi$	**B**$\Diamond$
2.	$\Diamond\Box\phi \to \Box\phi$	**S5**
3.	$\Box(\Diamond\Box\phi \to \Box\phi)$	2, **NEC**
4.	$\Box(\Diamond\Box\phi \to \Box\phi) \to (\Box\Diamond\Box\phi \to \Box\Box\phi)$	**K**
5.	$\Box\Diamond\Box\phi \to \Box\Box\phi$	3, 4, **MP**
6.	$\Box\phi \to \Box\Box\phi$	1, 5, **PL** (syllogism)

- So, as we have just proved, **S5** is strong enough to encompass both **B** and **S4**.

4.7 Soundness

- Next, for each system Δ (Δ = **K**, **D**, **T**, **B**, **S4** or **S5**), we now prove soundness.
  - Δ-soundness: every Δ-theorem is Δ-valid.

Soundness

- To prove soundness for all our systems, we would normally begin by proving the following theorem.

  Theorem I. If Γ is any set of modal wffs and $\mathfrak{M}$ is an MPL-model in which each wff in Γ is valid, then every theorem of **K**+Γ is valid in $\mathfrak{M}$.

- To make life easier for ourselves, we will assume that we have already proved this theorem—cf. Sider (2010, 173-174) for a proof.

- Given this proof, proving soundness for each system requires only that it is shown that all instances of the axiom schema associated with each system Δ is valid in every Δ-model.
- We will skip soundness of **K** as that follows trivially from **Theorem I**.

Soundness of T

- **T** is **K** + all instances of the **T**-axiom. So, what we need to show is that every theorem of **T** is valid in all **T**-models, i.e. valid in all reflexive models.

Proof.

— Assume for reductio that $\mathcal{V}_{\mathfrak{M}}(\Box\phi \rightarrow \phi, w) = 0$

— Hence, for some w, $\mathcal{V}_{\mathfrak{M}}(\Box\phi, w) = 1$ and $\mathcal{V}_{\mathfrak{M}}(\phi, w) = 0$.

— But since a **T**-model is reflexive, then $\mathcal{R}(w,w)$.

— Hence, if $\mathcal{V}_{\mathfrak{M}}(\Box\phi, w) = 1$ and $\mathcal{R}(w,w)$, then $\mathcal{V}_{\mathfrak{M}}(\phi, w) = 1$.

Contradiction

Soundness of S4

- **S4** is **K** + **T** + all instances of the **S4**-axiom. So, what we need to show is that every theorem of **S4** is valid in all **S4**-models, i.e. valid in all reflexive and transitive models.

Proof.

— Assume for reductio that $\mathcal{V}_{\mathfrak{M}}(\Box\phi \rightarrow \Box\Box\phi, w) = 0$

— Hence, for some w, $\mathcal{V}_{\mathfrak{M}}(\Box, w) = 1$ and $\mathcal{V}_{\mathfrak{M}}(\Box\Box\phi, w) = 0$

— If $\mathcal{V}_{\mathfrak{M}}(\Box\Box\phi, w) = 0$, then for some world w' such that $\mathcal{R}(w,w')$, $\mathcal{V}_{\mathfrak{M}}(\Box\phi, w') = 0$.

— But if $\mathcal{V}_{\mathfrak{M}}(\Box\phi, w') = 0$, then for some world w'' such that $\mathcal{R}(w,w'')$, $\mathcal{V}_{\mathfrak{M}}(\phi, w'') = 0$.

— However, since $\mathcal{V}_{\mathfrak{M}}(\Box\phi, w) = 1$, then for all worlds v such that $\mathcal{R}(w,v)$, $\mathcal{V}_{\mathfrak{M}}(\phi, v) = 1$.

— And since an **S4**-model is transitive, then if $\mathcal{R}(w,w')$ and $\mathcal{R}(w',w'')$, then $\mathcal{R}(w,w'')$.

— Hence, $\mathcal{V}_{\mathfrak{M}}(\phi, w'') = 1$

Contradiction

Completeness

- For some other time...

Chapter 5

Counterfactuals

5.1 Counterfactuals

- Counterfactuals are conditionals of the form in (11).

  (11) If it had been that ϕ, then it would have been that ψ.

- We will symbolize counterfactual conditionals using the symbol '$\Box\!\!\rightarrow$', i.e. we will symbolize the sentence in (12) as in (13).

  (12) If Oswald had not shot Kennedy, then somebody else would have.

  (13) $\phi \mathrel{\Box\!\!\rightarrow} \psi$

- Counterfactuals contrast indicative conditionals such as (14).

  (14) If Oswald did not shoot Kennedy, then somebody else did.

- Generally speaking, the characteristic mark of a counterfactual is subjunctive mood morphology (irrealis mood) and having an antecedent that is known (or at least *believed*) to be false.

5.2 The Behavior of Natural Language Counterfactuals

- Indicative conditionals are sometimes analyzed in terms of either material or strict conditionals.

$$\phi \rightarrow \psi \quad | \quad \phi ⥽ \psi$$

Remember, $\phi ⥽ \psi$ is defined as $\Box(\phi \rightarrow \psi)$

- However, there are good reasons to assume that neither of these analyses adequately captures the meaning of counterfactuals.

Contingently True Counterfactuals

- First, in contrast to indicative conditionals analyzed as strict conditionals, counterfactuals can be contingently true. For example, (12) seems true if (at the actual world) Oswald had a backup shooter.
- But that fails to make (12) necessarily true, hence our semantic for '$\Box\!\!\rightarrow$' should have the following features.

$$\begin{array}{rcl} P \Box\!\!\rightarrow Q & \nvDash & \Box(P \Box\!\!\rightarrow Q) \\ \neg(P \Box\!\!\rightarrow Q) & \nvDash & \Box\neg(P \Box\!\!\rightarrow Q) \end{array}$$

Antecedent Strengthening (Augmentation)

- Material conditionals and strict conditionals both license *antecedent strengthening* (or *augmentation*), i.e. the following inferences.

$$\frac{\phi \rightarrow \psi}{(\phi \wedge \chi) \rightarrow \psi} \qquad \frac{\phi \strictif \psi}{(\phi \wedge \chi) \rightarrow \psi}$$

- It is easy to demonstrate why these inferences are licensed — for material implication:

  · If '$\phi \rightarrow \psi$' is true, then either ϕ is false or ψ is true.
  — If ϕ is false, then $(\phi \wedge \chi)$ is also false.
  — If ψ is true, then the truth value of $(\phi \wedge \chi)$ is irrelevant.
  · So, regardless of the truth value of χ, '$(\phi \wedge \chi) \rightarrow \psi$' is true.

The same argument, relativized to worlds, carries directly over to the strict conditional.

- However, this inference pattern is not valid for counterfactuals. For example, the inference below is invalid.

(C1) If I had struck this match, it would have lit
(C2) Therefore, if I had struck this match and been in outer space, it would have lit.

- As a result, we need our semantics for counterfactuals to invalidate that inference, viz.

$$\phi \Box\!\!\rightarrow \psi \quad \nvDash \quad (\phi \wedge \chi) \Box\!\!\rightarrow \psi$$

No Contraposition

- Both the material and strict implication analysis licenses contraposition, i.e. the following inferences.

$$\frac{\phi \rightarrow \psi}{\neg\psi \rightarrow \neg\phi} \qquad \frac{\phi \strictif \psi}{\neg\psi \strictif \neg\phi}$$

- However, this inference pattern is invalid for counterfactuals.

(C3) If Herman had moved in, then Ernie would not have moved out.

(C4) Therefore, if Ernie had moved out, then Herman would not have moved in.

That this inference is invalid is easily demonstrated:

- Suppose that Ernie really wanted to live in the same house as Herman.
- However, Herman wanted to live in this house only because he thought it was a very nice house — regardless of whether Ernie lived there.
- Given these assumptions, (C3) is true while (C4) is false.

- As a result:

$$\phi \Box\!\!\rightarrow \psi \not\models \neg\psi \Box\!\!\rightarrow \neg\phi$$

Paradoxes of Material Implication

- The material implication analysis of conditionals licenses the following inferences:

$$\textbf{FA}\ \frac{\neg\phi}{\phi \rightarrow \psi} \qquad \textbf{TC}\ \frac{\psi}{\phi \rightarrow \psi}$$

- However, if these inferences were licensed for counterfactuals, the results would be very counterintuitive — for example, **FA** would entail that (almost) all counterfactuals are true!

- Hence, our semantics for counterfactuals should invalidate such inferences.

So, clearly, should a semantics for indicative conditionals, but we are setting that issue aside here.

$$\begin{array}{ccc} \neg P & \not\models & P \Box\!\!\rightarrow Q \\ Q & \not\models & P \Box\!\!\rightarrow Q \end{array}$$

Valid Inference Patterns

- Let's now consider some inference patterns that a semantics for counterfactuals *should* validate, for example the following.

$$\textbf{IR1}\ \frac{\phi \Box\!\!\rightarrow \psi}{\phi \rightarrow \psi} \qquad \textbf{IR2}\ \frac{\phi \strictif \psi}{\phi \Box\!\!\rightarrow \psi}$$

- That **IR1** is intuitively valid can be demonstrated as follows.
  — Suppose that '$\phi \Box\!\!\rightarrow \psi$' is true. If that counterfactual is true, it cannot intuitively be the case that 'ϕ' is true and 'ψ' is false.

- That **IR2** is intuitively valid can be demonstrated as follows.
  — If it is necessary that if ϕ then ψ, then surely if ϕ *had been true*, then ψ *would have been true* as well.

Context-Dependence

- The truth value of a counterfactual appears to depend on which facts are held fixed. For example.

  (15) If Edinburgh had been located in the Sahara desert, it would rain less in Edinburgh.

- There are (at least) two separate facts which can be held fixed here.

  (i) The location and span of the Sahara desert.

  (ii) The location and span of the city of Edinburgh.

- If (i) held fixed, then (15) seems true. But if (ii) is held fixed, (15) seems false.

5.3 The Lewis-Stalnaker Theory

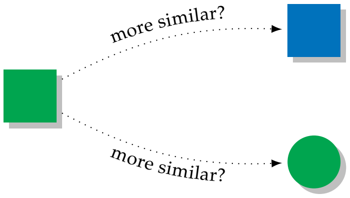
- The general idea behind the Lewis-Stalnaker theory of counterfactuals is the following.

  > A counterfactual, $\phi \mathrel{\Box\!\!\rightarrow} \psi$, is true if and only if at the worlds where ϕ is true, the world most similar to the actual world (the world of evaluation) is a world where ψ is true.

- Lewis uses the following example to illustrate.

  (16) If kangaroos had no tails, they would topple over.

- (16) seems true, however it would clearly be false if the world at which the counterfactual was evaluated was a world where kangaroos had crutches or wings instead of tails, or if it was a world where the laws of gravity were different.
- Hence, what we *actually* do when evaluating counterfactuals is look for the world *most* similar to the actual world where the antecedent is true, and then check the truth value of the consequent there.
- The key notion here is of course that of *similarity*.
- As mentioned above, counterfactuals seem context-sensitive. According to Lewis and Stalnaker, this context-sensitivity can be captured in terms of the similarity relation, i.e. which similarity relation is relevant is context-sensitive.
- For example, consider the figure below. The green square seems more similar to the blue square with regards to shape, but more similar to the circle with respect to color.



- As regards the question 'which is more similar?' — there is clearly no correct answer. So, with respect similarity, it depends on the context which factors are to be considered relevant.
- And so, as regards (15), we could specify the similarity relation in (at least) two ways.
  - Specify similarity in terms of the actual span of the Sahara desert. In this case, the actual location of Edinburgh would have to be changed to make the antecedent true.
  - Specify similarity in terms of the actual location of Edinburgh. In this case, the span of the Sahara desert would have to be expanded so as to include Edinburgh in its present location.

5.4 Stalnaker's System

- Let's refer to Stalnaker's system as **SC**.

5.4.1 Primitive Vocabulary of SC

- The modal language SC contains the following expression types.
  - All expressions of MPL + the connective '$\Box\!\!\rightarrow$'.

We will generally assume an S5-modal system to make things easier.

5.4.2 Syntax of SC

- The syntactic rules of SC, are as follows
  - Sentence letters are wffs.
  - If ϕ and ψ are wffs, then $(\phi \rightarrow \psi)$, $\neg\phi$, $\Box\phi$, and $(\phi \Box\!\!\rightarrow \psi)$ are wffs.
  - Nothing else is a wff.

The remaining connectives are defined in the usual way.

5.4.3 Semantics and Models for SC

- To give a semantics for counterfactuals, we first need to define the notion of an **SC**-model.
  - An **SC**-model, $\mathfrak{M}^{\text{SC}}$, is an ordered triple $\langle \mathcal{W}, \preceq, \mathcal{I} \rangle$ where:
    - $\mathcal{W}$ is a a non-empty set (of worlds).
    - $\mathcal{I}$ is a two-place function: $\mathcal{I}$: $S{\times}\mathcal{W} \longmapsto \{0,1\}$ (interpretation function)
    - $\preceq$ is a three-place relation over $\mathcal{W}$ (nearness relation)
    - The valuation function $\mathcal{V}_{\mathfrak{M}}$ for $\mathfrak{M}$ and $\preceq$ satisfy the following conditions:
      1. For any $w \in \mathcal{W}$: $\preceq_w$ is *strongly connected* in $\mathcal{W}$.
      2. For any $w \in \mathcal{W}$: $\preceq_w$ is *transitive*.
      3. For any $w \in \mathcal{W}$: $\preceq_w$ is *anti-symmetric*.
      4. For any $x, y \in \mathcal{W}$: $x \preceq_x y$ (base)

5. For any SC-wff ϕ, provided that $\mathcal{V}_{\mathfrak{M}}(\phi,v) = 1$ for some $v \in \mathcal{W}$:
   — for every $z \in \mathcal{W}$, there is some $w \in \mathcal{W}$ such that $\mathcal{V}_{\mathfrak{M}}(\phi,w) = 1$ and such that for any $x \in \mathcal{W}$, if $\mathcal{V}_{\mathfrak{M}}(\phi,x) = 1$, then $w \leq_z x$. (limit)

- Some explanations:
  - '$\boldsymbol{x \leq_z y}$' is a binary relation (relative to z) that holds between x and y. We will sometimes write this as $\mathcal{R}_z(x, y)$.
  - A binary relation $\mathcal{R}$ is **strongly connected** in a set A if and only if for each $x, y \in A$, either $\mathcal{R}(x, y)$ or $\mathcal{R}(y, x)$.
  - A binary relation is **anti-symmetric** in a set A if and only if for all $x, y \in A$, $x = y$ whenever $\mathcal{R}(x, y)$ and $\mathcal{R}(y, x)$.
  - So, if $x \leq_z y$ and $y \leq_z x$, then $x = y$.
- Next, we need to define the notion of a valuation – we'll focus here only on the clause for counterfactuals.
  - Where $\mathfrak{M} = \langle \mathcal{W}, \leq, \mathcal{I} \rangle$ and $\mathfrak{M}$ is any **SC**-model, the **SC**-valuation for $\mathfrak{M}$, $\mathcal{V}_{\mathfrak{M}}$, is defined as the two-place function that asssign either 0 or 1 to each **SC**-wff relative to each member of $\mathcal{W}$, subject to the following constraints: Where α is any sentence letter, ϕ and ψ are any wffs, and w is any member of $\mathcal{W}$:

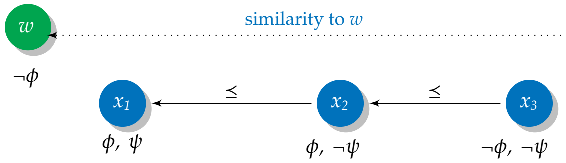
$\mathcal{V}_{\mathfrak{M}}(\phi \mathrel{\Box\!\!\rightarrow} \psi, w)$ = 1 iff for any $x \in \mathcal{W}$,
IF $\mathcal{V}_{\mathfrak{M}}(\phi, x) = 1$ and for any $y \in \mathcal{W}$ such
that $\mathcal{V}_{\mathfrak{M}}(\phi, y) = 1$, $x \leq_w y$]
THEN: $\mathcal{V}_{\mathfrak{M}}(\psi, x) = 1$

- **The Nearness (Similarity) Relation**: $\leq$
  - The key notion in this semantics is clearly the three-place relation $\leq$.
  - The meaning of the statement '$x \leq_w y$' is simply that relative to the world w, x is as least as near (or similar) to w as y.
  - Hence, the valuation clause for '$\phi \mathrel{\Box\!\!\rightarrow} \psi$' then says that **IF the following holds**: If ϕ is true at x and for all other worlds y where ϕ is also true, x is at least as near (or similar) to w as y, then ψ is true at x, **THEN the counterfactual '$\phi \mathrel{\Box\!\!\rightarrow} \psi$' is true** (and false otherwise).

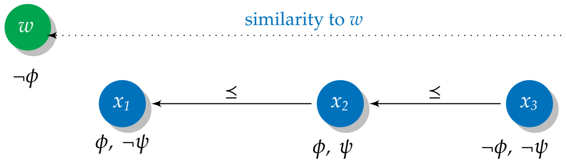
- **Constraints on the Nearness (Similarity) Relation**: $\leq$
  - **Strong Connectivity**
    Because the relation $\leq$ is strongly connected, any two worlds x and y can be compared for similarity relative to any given world w.

- **Anti-Symmetry**
  Because the relation $\leq$ is anti-symmetric, no two worlds will ever be equally near (equally similar) to any given world w.
- **Base**
  The 'base' assumption ensures that every world x is at least as similar to itself as it is to any other world — hence given anti-symmetry every world must be more similar to itself than any other (distinct) world.
- **Limit**
  The 'limit' assumption guarantees that there always is a unique set of closest worlds. I.e. an infinite chain of ever more similar worlds is ruled out. Combined with anti-symmetry, it follows that there always is one unique closest world.

SC-model 1: $\mathcal{V}_{\mathfrak{M}}(\phi \;\Box\!\!\rightarrow \psi, w) = 1$



SC-model 2: $\mathcal{V}_{\mathfrak{M}}(\phi \;\Box\!\!\rightarrow \psi, w) = 0$



5.4.4 Semantic Validity Proofs in **SC**

- Since we now have a fully functional semantics for '$\Box\!\!\rightarrow$', we can do semantic validity proofs, cf. below. (in the proof, assume that $\mathfrak{M}$ is an SC-model).

  Proof. $\vDash_{SC} (P \wedge Q) \rightarrow (P \Box\!\!\rightarrow Q)$

— Suppose for reductio that $\mathcal{V}_{\mathfrak{M}}((P \wedge Q) \rightarrow (P \Box\!\!\rightarrow Q), w) = 0$

— If so, $\mathcal{V}_{\mathfrak{M}}(P \wedge Q, w) = 1$.

— And, $\mathcal{V}_{\mathfrak{M}}(P \Box\!\!\rightarrow Q, w) = 0$.

— If $\mathcal{V}_{\mathfrak{M}}(P \Box\!\!\rightarrow Q, w) = 0$, then at the closest P-world to w, Q is false. So,

i. there is a world v such that $\mathcal{V}_{\mathfrak{M}}(P, v) = 1$.

ii. for all worlds u, if $\mathcal{V}_{\mathfrak{M}}(P, u) = 1$, then $v \preceq_w u$.

iii. $\mathcal{V}_{\mathfrak{M}}(Q, v) = 0$.

— From $\mathcal{V}_{\mathfrak{M}}(P \wedge Q, w) = 1$, it follows that $\mathcal{V}_{\mathfrak{M}}(P, w) = 1$.

— Since $\mathcal{V}_{\mathfrak{M}}(P, w) = 1$, $v \preceq_w w$.

— But by (**base**) $w \preceq_w v$, and so since '$\preceq$' is anti-symmetric, $v = w$.

— From $\mathcal{V}_{\mathfrak{M}}(P \wedge Q, w) = 1$, it follows that $\mathcal{V}_{\mathfrak{M}}(Q, w) = 1$.

— But since $v = w$ and $\mathcal{V}_{\mathfrak{M}}(Q, v) = 0$, it follows that $\mathcal{V}_{\mathfrak{M}}(Q, w) = 0$.

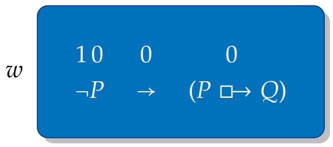
Contradiction

5.4.5 Semantic Invalidity in **SC**

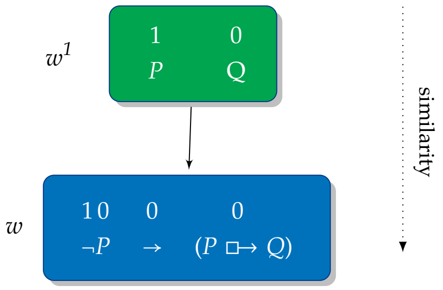
- To demonstrate invalidity in **SC**, we once again use countermodels. These countermodels will now look slightly different.

    - Worlds are represented as rounded boxes, but arrows now indicate the nearness (similarity) relation.

- Let's construct a countermodel for the formula '$\neg P \rightarrow (P \mathrel{\Box\!\!\rightarrow} Q)$'. First, we enter the formula, the invalidity to be established, and "forced" truth values.



- Next, we need to evalute the counterfactual. This requires looking at the nearest possible worlds.



- $\neg P$ is true at w.
- But, $(P \mathrel{\Box\!\!\rightarrow} Q)$ is false at w, because at the nearest P-world, namely w^1, Q is false.

- THE OFFICIAL MODEL

$\mathcal{W}$	w, w^1
$R_{\leq w}$	$\langle w, w^1 \rangle$
$\mathcal{I}$	w : $Q, \neg P$ w^1 : $P, \neg Q$

5.4.6 Logical Features of **SC**

- Let's check how Stalnaker's proposed analysis of counterfactuals fares with respect to the desiderata mentioned earlier.

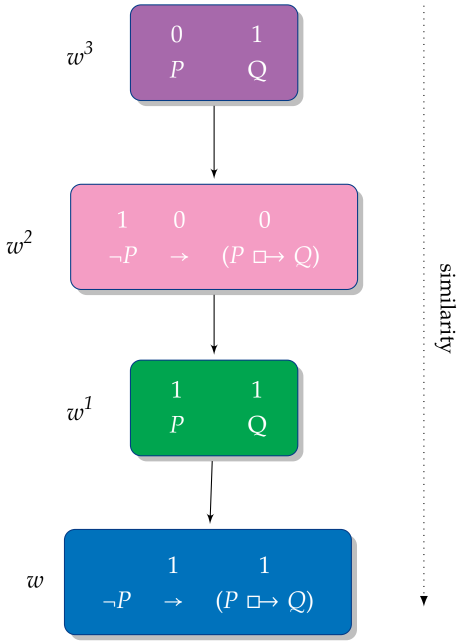
    1. We observed that counterfactuals are not intuitively true simply because their antecedents are false or because their consequents are true. We have already proved **SC1** above and it would be easy to prove **SC2** as well.

        SC1 $\neg P \nvDash_{\mathbf{sc}} P \mathrel{\Box\!\!\rightarrow} Q$

SC2 $Q \nvDash_{\mathbf{sc}} P \,\Box\!\!\rightarrow Q$

Homework exercise: Construct a counter-model proving the invalidity in **SC2**.

2. We observed that counterfactuals can be contingently true, hence: $P \,\Box\!\!\rightarrow Q \nvDash_{\mathbf{sc}} \Box(P \,\Box\!\!\rightarrow Q)$. Stalnaker's analysis delivers this result — consider, for example, the following model.

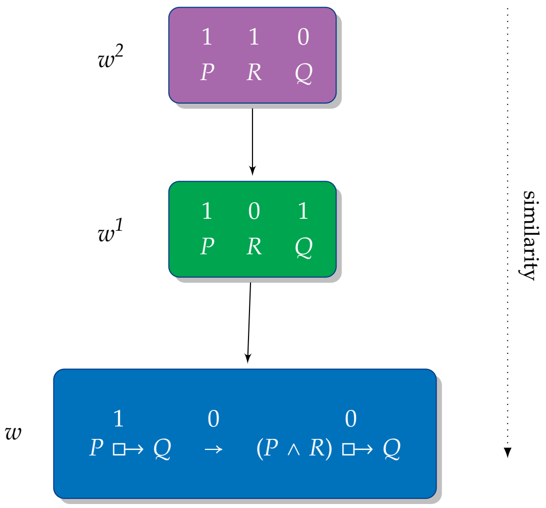

w^3
0 1
P Q
w^2
1 0 0
$\neg P$ → $(P \,\Box\!\!\rightarrow Q)$
w^1
1 1
P Q
w
1 1
$\neg P$ → $(P \,\Box\!\!\rightarrow Q)$
similarity

In this model, call it $\mathfrak{M}^{\mathbf{sc}}$, we have (a) and (b).

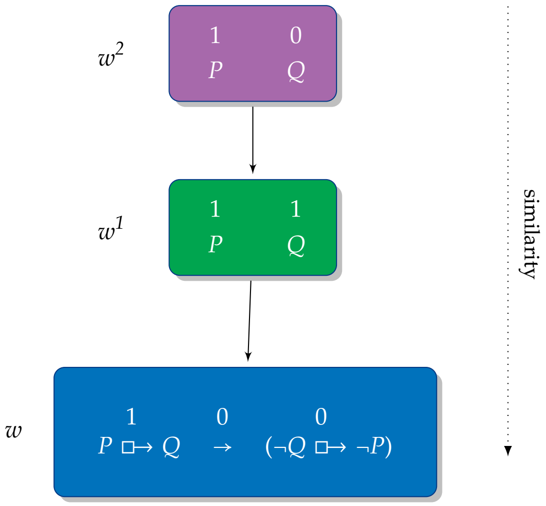
(a) $\mathcal{V}_{\mathfrak{M}^{\mathrm{sc}}}(P \,\Box\!\!\rightarrow Q, w) = 1$ **(b)** $\mathcal{V}_{\mathfrak{M}^{\mathrm{sc}}}(P \,\Box\!\!\rightarrow Q, w^2) = 0$

Hence, it follows that: $P \,\Box\!\!\rightarrow Q \nvDash_{\mathfrak{M}^{\mathrm{sc}}} \Box(P \,\Box\!\!\rightarrow Q)$

3. Third, we saw that counterfactuals do not license augmentation (antecedent strengthening). Hence, $P \,\Box\!\!\rightarrow Q \nvDash_{\mathbf{sc}} (P \wedge R) \,\Box\!\!\rightarrow Q$ — and again Stalnaker's system delivers this result.


w^2
1 1 0
P R Q
w^1
1 0 1
P R Q
w
1 0 0
$P \mathrel{\Box\!\!\rightarrow} Q \rightarrow (P \wedge R) \mathrel{\Box\!\!\rightarrow} Q$
similarity

4. Contraposition also fails for counterfactuals. And again, Stalnaker's system delivers this result.


w^2
1 0
P Q
w^1
1 1
P Q
w
1 0 0
$P \mathrel{\Box\!\!\rightarrow} Q \rightarrow (\neg Q \mathrel{\Box\!\!\rightarrow} \neg P)$
similarity

- Other logical features that Stalnaker's system correctly captures include:
  - **No importation**: $\phi \mathrel{\Box\!\!\rightarrow} (\psi \mathrel{\Box\!\!\rightarrow} \chi) \nvDash_{\textbf{sc}} (\phi \wedge \psi) \mathrel{\Box\!\!\rightarrow} \chi$
  - **No permutation**: $\phi \mathrel{\Box\!\!\rightarrow} (\psi \mathrel{\Box\!\!\rightarrow} \chi) \nvDash_{\textbf{sc}} \psi \mathrel{\Box\!\!\rightarrow} (\phi \mathrel{\Box\!\!\rightarrow} \chi)$
  - **No transitivity**: $\{\phi \mathrel{\Box\!\!\rightarrow} \psi, \psi \mathrel{\Box\!\!\rightarrow} \chi\} \nvDash_{\textbf{sc}} \phi \mathrel{\Box\!\!\rightarrow} \chi$
- Check Sider (2010, 218) for proofs.

5.5 Lewis Criticism of Stalnaker's System

- Lewis' analysis of counterfactuals is also a similarity based analysis, however Lewis objects to two specific aspects of Stalnaker's proposed analysis, namely the anti-symmetry assumption and the limit assumption.
- Remember:
  - **The Limit Assumption**
    Guarantees that there always is a set of most similar worlds.
  - **The Anti-Symmetry Assumption**
    Guarantees that there are no ties in similarity among worlds.
  - Jointly, these assumptions entail uniqueness.
  - **The Uniqueness Assumption**
    There is always a unique most similar world.

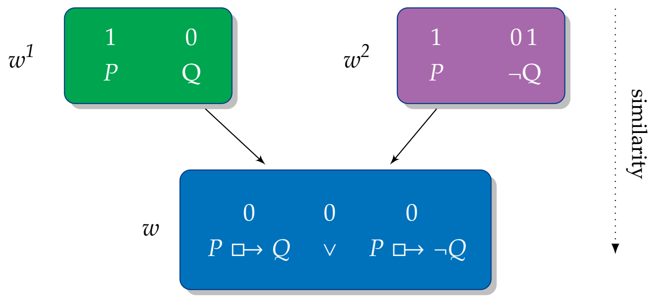
Against Anti-Symmetry

- A couple of important schemas are valid only if anti-symmetry is assumed, e.g.

CEM $(\phi \mathrel{\Box\!\!\rightarrow} \psi) \vee (\phi \mathrel{\Box\!\!\rightarrow} \neg\psi)$ (Conditional Excluded Middle)

DIST $\big(\phi \mathrel{\Box\!\!\rightarrow} (\psi \vee \chi)\big) \rightarrow \big((\phi \mathrel{\Box\!\!\rightarrow} \psi) \vee (\phi \mathrel{\Box\!\!\rightarrow} \chi)\big)$ (Distribution)

- If anti-symmetry is given up, then it is easy to see that there are counterexamples to e.g. **CEM**.


w^1
1 0
P Q
w^2
1 01
P ¬Q
w
0 0 0
$P \mathrel{\Box\!\!\rightarrow} Q$ ∨ $P \mathrel{\Box\!\!\rightarrow} \neg Q$
similarity

- Notice that in this countermodel, the following holds:
  - $w^1 \preceq_w w^2$

- $w^2 \leq_w w^1$
- $w^1 \neq w^2$ (No Anti-Symmetry)

- Whether **CEM** or **DIST** ought to be given up are substantial philosophical questions.
- Both principles have a great deal of intuitive plausibility. Consider **CEM**:
- The following is an equivalent formulation of **CEM**:

$$\neg(\phi \mathrel{\Box\!\!\rightarrow} \psi) \rightarrow (\phi \mathrel{\Box\!\!\rightarrow} \neg\psi)$$

  - Whenever ϕ is possibly true, everyone agrees that the converse of the formula above is true.

$$(\phi \mathrel{\Box\!\!\rightarrow} \neg\psi) \rightarrow \neg(\phi \mathrel{\Box\!\!\rightarrow} \psi)$$

  - I.e. Suppose we assume that if Mary had come to the party, it would not have been fun. From this it intuitively follows that it is not the case that if Mary had come to the party, it would have been fun.
  - But, intuitively at least, it is hard to tell these two formulas apart, i.e. (17) and (18) are equivalent.

(17) If Mary had come to the party, it would not have been fun.

(18) It's not the case that if Mary had come to the party, it would have been fun.

  - But if these are truth conditionally equivalent, and everyone agrees that the converse is true, **CEM** is true.

- Lewis' provides several arguments against **CEM** and **DIST**.
- **First**, Lewis makes the quite reasonable observation that ties between worlds just seem clearly possible — For example:
  - Suppose there are two worlds w^1 and w^2.
  - Suppose at w^1 my legs are 0.01 inches shorter than they are at w and suppose at w^2 my legs are 0.01 inches longer than they are at w.
  - Finally, suppose that w^1 and w^2 are identical to w in every other respect.
- It is hard to accept that either w^1 or w^2 is more similar to w than the other.

- **Second**, Lewis argues that intuitions are less clear here than typically assumed. Consider for example the following:
  - It's not the case that if I had flipped this coin it would have landed heads.
  - It's also not the case that if I had flipped this coin it would have landed tails.
  - However, it is the case that if I had flipped this coin, it would have landed either heads or tails.

- Rather, what one should say about these cases is:
  - If I had flipped this coin, it *might* have landed heads, but it's not the case that it *would* have landed heads
- Lewis' analyzes such *might*-counterfactuals, i.e. statements of the form "if it had been that ϕ, then it might have been that ψ" as:

$$\neg(\phi \mathrel{\Box\!\!\rightarrow} \neg\psi)$$

- Such might-counterfactuals are intuitively weaker than the standard *would*-counterfactual, i.e. statements of the form "if it had been that ϕ, then it would have been that ψ" — symbolized by:

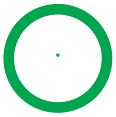
$$\phi \mathrel{\Box\!\!\rightarrow} \psi$$

- But if **CEM** is valid, *might*-counterfactuals entail *would*-counterfactuals, i.e.

$$\neg(\phi \mathrel{\Box\!\!\rightarrow} \neg\psi) \rightarrow (\phi \mathrel{\Box\!\!\rightarrow} \psi)$$

Against Limit

- Consider the circle below.



- The following counterfactual is false:

(19) If the radius of the circle had been more than one inch long, it would have been 1.000 inches long.

- This counterfactual is predicted to be true by Stalnaker's analysis, since there appears to be no most similar world where the radius of the circle is more than one inch long.
- For every world w' where the radius of the circle is $1+k$ inches long, there is another world w'' where the radius of the circle is $1+\frac{k}{2}$ inches long.
- Hence, the counterfactual is predicted to be trivially true — but that seems to be the wrong prediction.

5.5.1 Lewis' System (LC)

- LC-models and their valuation functions ($\mathcal{LV}$) are like Stalnaker's except for the following differences:
  - **Anti-Symmetry** and **Limit** are not assumed.
  - **Base** is changed to: for any x and y, if $y \leq_x x$, then $x = y$.
  - The truth condition for '$\Box\!\!\rightarrow$' is changed to the following:

$\mathcal{LV}_{\mathfrak{M}}(\phi \mathrel{\Box\!\!\rightarrow} \psi, w) = 1$ iff

EITHER: ϕ is true in no worlds,

OR: there is some world, x, such that $\mathcal{LV}^{\mathfrak{M}}(\phi, x) = 1$ and for all y, if $y \leq_w x$, then $\mathcal{LV}^{\mathfrak{M}}(\phi \rightarrow \psi, y) = 1$.

- Why the new semantics for '$\Box\!\!\rightarrow$'?
  - The limit assumption is now allowed to fail and Lewis' revised analysis of '$\Box\!\!\rightarrow$' is designed to capture cases that Stalnaker got intuitively incorrect results for (vacuously true counterfactuals do to a failure of the limit assumption).
  - Think about the case involving the circle above—which Stalnaker predicted to be vacuously true.
  - On Lewis' view, (if ϕ is possibly true), '$\phi \mathrel{\Box\!\!\rightarrow} \psi$' is true iff there is a ϕ-world with the following feature:
    - — No matter how much closer in similarity to w you go, you'll never find a world where ϕ is true and ψ is false.
  - This now ensures that (19) comes out false.

(19) If the radius of the circle had been more than one inch long, it would have been 1.000 inches long.

  - The only way for this to be true is for there to be a world w' where the antecedent is true and where for each world w'' which is as similar to w as w' (or more similar), the antecedent is never true at w' while the consequent is false at w'.
  - But it's quite clear that this condition is not satisfied. Let w' be a world where antecedent and consequent are both true. Is there a world w'' which is more similar to w than w' where the antecedent is true and the consequent is false? Clearly yes, e.g. a world where the radius of the circle is 2 inches long.

5.6 Disjunctive Antecedents: Problems for Stalnaker and Lewis

- Natural language appears to validate the following inference pattern:

$$(\phi \vee \psi) \mathrel{\Box\!\!\rightarrow} \chi \quad \Rightarrow \quad \phi \mathrel{\Box\!\!\rightarrow} \chi$$

- For example,

(20) If I had offered compensation or had apologized, I would have admitted guilt.

- (20) seems to license the following inference.

(21) If I had offered compensation, I would have admitted guilt.

- Notice that there is something quite odd about this — normally disjunctions do not allow inferences to either disjunct, i.e. the following are clearly invalid inferences.

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

- But in counterfactuals (and conditionals more generally), inferences similar to this kind are apparently acceptable.
- Yet, neither Stalnaker nor Lewis can capture that if '$(\phi \vee \psi) \mathrel{\Box\!\!\rightarrow} \chi$' is true, then '$\phi \mathrel{\Box\!\!\rightarrow} \chi$' is true as well.

A Possible Response: A General Problem with Disjunctions

- In complex environments, disjunctions tend to behave in quite notorious ways — for example (22) licenses both inferences in (22a) and (22b).

  (22) You may travel by car or by boat.
  a. ⇒ You may travel by car.
  b. ⇒ You may travel by boat.

- Hence, this is really a problem about disjunction and not a problem specific to Stalnaker or Lewis' analyses.

Chapter 6

Decision Theory

6.1 Decision and Game Theory

- Crudely speaking, decision theory is the study of decision problems and optimal decision making.
- The simplest form that such a decision problem can take is the following:

	State 1	*State 2*
Choice 1	a	b
Choice 2	c	d

- This decision table breaks down as follows:
  - **Choices** are actions or options that an agent has available to her.
  - **States** are the ways the world could turn out which affects the value of the outcome of the agent's choice.
- Suppose you have a choice between watching your favorite football team play a game or work on a paper due in a few days. Suppose further that each of the possible outcomes are ranked as follows.

	Team Wins	*Team Loses*
Watch Football	4	1
Work on Paper	2	3

- From this ordering it follows that:
  - If your team wins, then watching the game is the best decision and if your team loses, then working on the paper is the best decision.

  - However, watching your team lose is clearly the worst scenario, whereas working on your paper while your team wins is a slightly less bad scenario.

- Sometimes the states relevant to your decision are actions of another agent, i.e. in a game.

	Rock	*Paper*	*Scissor*
Rock	0,0	-1,1	1,-1
Paper	1,-1	0,0	-1,1
Scissor	-1,1	1,-1	0,0

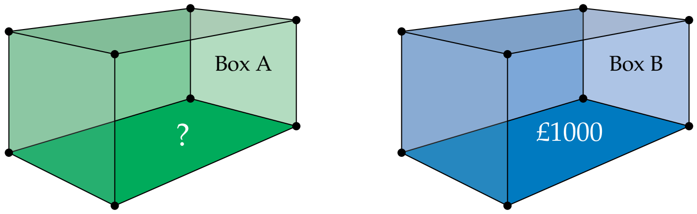
- The decision table now has two numbers—the first presenting how good the outcome is for you, the second how good the outcome is for your opponent.
- One typically draws a sharp distinction between decision problems that depend only on the impersonal world and decision problems that depend on the actions of others. The former are standardly said to be problems in *decision theory* whereas the second are said to be problems in *game theory*.

One thing that is highly relevant to making an optimal decision is obviously knowing the likelihood of each of the states obtaining, i.e. the probability. We'll look at probability theory later.

6.1.1 Some (famous) Decision Problems

Newcomb's Problem

- In front of you are two boxes **A** and **B**. You can see that **B** contains £1000, but you cannot see what is in **A**.



- You have two options:

  O1 Take **A**.

  O2 Take both **A** and **B** including the £1000.

- However, there is a catch (obviously).
  - A demon has predicted whether you will take just one box or two boxes. The demon is very good at predicting these things and has in the past made many predictions all of which have turned out to be correct.
  - If the demon predicts that you'll take both boxes, then she puts nothing in **A**. However, if the demon predicts that you'll take just one box, she'll put £1,000,000 in **A**.

- Hence, we have the following decision table.

	Predict 1 Box	*Predict 2 Boxes*
Take 1 Box	£1,000,000	£0
Take 2 Boxes	£1,001,000	£1,000

- The problem now is that there are compelling argument for both decisions.

Take 1 Box
Every time the demon has made the prediction that the person would take just one box, the person won a £1,000,000. So, take 1 box and leave with a million pounds.

Take 2 Boxes
Either the demon has put £1,000,000 in **A** or she has not. Either way, taking two boxes yields the better outcome: If she has, you win £1,001,000 rather than £1,000,000 and if she has not, you win £1,000 rather than £0.

- So, what is the correct decision?

Voting

- Suppose that 9 members of a committee are trying to decide between three candidates *A*, *B*, and *C*.

  - 4 members rank the candidates as follows: $A > B > C$
  - 3 members rank the candidates as follows: $C > B > A$
  - 2 members rank the candidates as follows: $B > C > A$

- Now depending on the reasoning, a case can be made for either candidate:

Argument for A
A has the most votes, so *A* should get the job.

Argument for B
A majority rank *B* over *A*, and a majority also rank *B* over *C*. So, one could argue (as many have) that when one candidate is preferred by a majority over each of their rivals, they should win.

Argument for C
No candidate has a majority, so the two candidates with most votes should have a run-off against eachother. *B* has the least votes so will be ruled out, but the people who preferred *B*, prefer *C* to *A*, so they'll change their votes to *C*. If so, *C* will have more votes than *A*.

- So, what is the correct decision?

Decision Guides: Dominance Reasoning

- One of the simplest (and least controversial) principles in decision theory is *the prohibition against choosing dominated options*.

6.2 Dominance

- There are two versions of dominance, namely a weak and a strong version.

  - **Strong Dominance**
    An option *A strongly dominates* an option *B* if—in every state—*A* leads to a better outcome than *B*.

  - **Weak Dominance**
    An option *A weakly dominates* an option *B* if—in every state—*A* leads to an outcome at least as good as *B* and in some state(s) leads to a better outcome than *B*.

- Suppose your choices are ranked as follows relative to the possible states.

		Team Wins	*Team Loses*
A1	**Watch Football**	2	1
A2	**Work on Paper**	4	3

- Since 4 > 2 and 3 > 1, **A2** *strongly dominates* **A1**. Hence, you should work on the paper.
- Notice that **Weak Dominance** is a stronger principle than **Strong Dominance**. In the example above, working on the paper also *weakly dominates* watching football.

6.3 States, Choices, and Independence

- One tacit, but extremely important, assumption that we have made so far, is that our choices are independent of the states of the world.
- Without this assumption, dominance reasoning yields some rather unfortunate results. Consider the following example:

Joyce (1999) credits this observation to Jeffrey.

> Suppose you have just parked in a seedy neighborhood when a man approaches and offers to "protect" your car from harm for $10. You recognize this as extortion and have heard that people who refuse "protection" invariably return to find their windshields smashed. Those who pay find their cars intact. You cannot park anywhere else because you are late for an important meeting. It costs $400 to replace a windshield. Should you buy "protection"? **Dominance** says that you should not. Since you would rather have the extra $10 both in the even that your windshield is smashed and in the event that it is not, **Dominance** tells you not to pay.
>
> Joyce (1999, 115-116)

- Here is the decision table.

		Windshield Broken	*Windshield Intact*
A1	**Pay Insurance**	–$410	–$10
A2	**Don't Pay Insurance**	–$400	$0

- **A2** strongly dominates **A1**, so dominance recommends that you do not pay insurance. Yet, this seems to be clearly wrong.
- So what's the problem? The problem is that your choices have direct effects on the states of the world!
- To avoid this problem, we will assume for now that the states obtaining are independent of choices.
- Later, we will discuss the notion of independence in more detail.

6.4 Maximax and Maximin

- Two other decision principles, **Maximax** and **Maximin**.
  - **Maximax**
    Always maximise the maximum outcome possible.
  - **Maximin**
    Always maximise the minimum outcome possible.
- So, consider the following decision table.

		Team Wins	*Team Loses*
A1	**Watch Football**	4	1
A2	**Work on Paper**	2	3

  - Using Maximax as your decision rule, you should perform action **A1**. The reason is that you can only get the best possible outcome, (4), by watching football.
  - Using Maximin as your decision rule, you should perform action **A2**. The reason is that of the worst possible potential outcomes in each state, **A2** guarantees the least bad.
- Note that the **Maximax** rule has some rather disconcerting consequences. For example, if you were thinking about investing money, the **Maximax** rule would recommend investing in whatever had the highest potential payoff regardless of the chances of that outcome.
- So, suppose you decided to bet on the winner of the Champions League, then **Maximax** would recommend betting all your money on Celtic (even if, say, Real Madrid was paying only slightly less than Celtic).

6.5 Ordinal vs. Cardinal Utilities

- So far we've only focused on what is standardly called *ordinal* utilities, i.e. an ordering of the available outcomes. So, if O_n is an outcome, we simply have a ordering of the form $O_1 > O_2 > O_3 > O_4$.
- In other words, we have not taken into account *how much* we might prefer one outcome over another.
- To see why ordinal utilities are often inadequate consider this problem:

> Suppose you have a choice between two airlines, **Airline 1** and **Airline 2**. **Airline 1** is cheap but less reliable than **Airline 2**. Let's suppose that **Airline 1** runs great in good weather, but not so great in bad weather — and also that you have no way of determining what the weather will be like. You would prefer saving money but also prefer things not going badly wrong. So, we have the following decision table.

	Good Weather	*Bad Weather*
Airline 1	4	1
Airline 2	3	2

- But now consider two possible further specifications of the problem:

$\boxed{\text{S1}}$ When the weather is bad, **Airline 1**'s luggage workers work really slow. So, your luggage could be delayed by one or more days.

$\boxed{\text{S2}}$ When the weather is bad, **Airline 1**'s planes tend to crash.

- The difference between these ways of further specifying the decision problem is not reflected in the decision table above.
- Clearly, if S1 is the case, it might be worth going with the cheaper airline, **Airline 1**.
- However, if S2 is the case, it would clearly NOT be worth flying **Airline 1** to save money since you would effectively be risking your life!
- In short, **decision tables should be sensitive to the magnitude of difference between the possible outcomes**.
- When the numbers assigned to outcomes reflect how much we prefer one option over another we are using what is called *cardinal utilities*.
- Using cardinal utilities, the decision table above could be specified as follows:

Homework Exercise: Determine which actions **Maximax** and **Maximin** would recommend for each of these decision tables.

S1	*Good Weather*	*Bad Weather*
Airline 1	20	4
Airline 2	6	5

S2	*Good Weather*	*Bad Weather*
Airline 1	20	-1000
Airline 2	6	5

The Regret Strategy

- One strategy for choosing between different outcomes that are assigned various cardinal utilities is to use the notion of *regret*.
- Consider the following decision table.

Table 1		*Sunny*	*Rain*	*Thunderstorm*
A1	**Picnic**	20	5	0
A2	**Football**	15	2	6
A3	**Movies**	8	10	9

- We define regret in terms of the difference between the value of the best choice and the other choices given the state of the world. Hence, the regret table associated with Table 1 is Regret 1 below.

Regret 1		*Sunny*	*Rain*	*Thunderstorm*
A1	**Picnic**	0	5 (10–5)	9 (9–0)
A2	**Football**	5 (20–15)	8 (10–2)	3 (9–6)
A3	**Movies**	12 (20–8)	0	0

- With the notion of regret, we can describe another range of decision principles, e.g. **Minimax Regret**.
  - **Minimax Regret**
    Choose the outcome with the lowest maximum possible regret.
- So, in this case, **Minimax Regret** recommends playing football over having a picnic or going to the movies.

Regret 1		*Sunny*	*Rain*	*Thunderstorm*
A1	**Picnic**	0	5	9
A2	**Football**	5	8	3
A3	**Movies**	12	0	0

- However, **Minimax Regret** does have some rather odd consequences.
- Using the **Minimax Regret** rule, going to the movies is deemed the worst option of all (since this outcome is associated with highest possible regret, 12).
- But now suppose that, for some reason, having a picnic is ruled out (suppose picnics are illegal). Since that option is ruled out, Table 2 is now the relevant decision table.

Table 2		*Sunny*	*Rain*	*Thunderstorm*
A2	**Football**	15	2	6
A3	**Movies**	8	10	9

- And the regret table for Table 2 is then Regret 2 below and the maximum possible regret is now associated with playing football.

Regret 2		*Sunny*	*Rain*	*Thunderstorm*
A2	**Football**	0	8	3
A3	**Movies**	7	0	0

- So, from ruling out what before was an irrelevant option, namely having a picnic, the recommendation made by **Minimax Regret** suddenly changes. That seems quite strange.

This is also called the *contraction condition*, *(Amartya) Sen's alpha property*, or *Chernoff's condition*.

- In short, **Minimax Regret** violates a principle called *Irrelevance of Independent Alternatives*.
  - **Irrelevance of Independent Alternatives**
    If an option C is chosen from some set of options S, then C should be chosen from any set of options T where (*a*) $C \in T$ and (*b*) $T \subseteq S$.

6.6 Do What Is Likely To Work

- Every day decisions intuitively involve some estimate of the likelihood of various states of the world obtaining.
- Hence, a rule that it might be natural to suppose one should follow is the rule of doing what is *likely* to work best.
  - To give an example, in the decision table below, if it is assumed that sunshine is more likely than rain or thunderstorms, then we should have a picnic.

Table 1		*Sunny*	*Rain*	*Thunderstorm*
A1	**Picnic**	20	5	0
A2	**Football**	15	2	6
A3	**Movies**	8	10	9

  - Similarly, if we believe that rain or thunderstorms are the most likely, we should go to the movies.

- One neat thing about this principle is that it only requires *ordinal* ranking of likelihoods.
- I.e. all that is needed is a ranking of each state of the world in terms of how likely they are, viz. $\Pr(\mathit{Sunny}) > \Pr(\mathit{Rain}) > \Pr(\mathit{Thunderstorm})$. (we do not need to assign numerical probabilities to each of the states).
- Sadly, this approach also has some rather unfortunate consequences. Suppose you're facing the following decision:

Table 3		*Infection*	*No Infection*
A1	**Vaccine**	500	-10
A2	**No Vaccine**	-5000	10

- Suppose you have been exposed to a deadly virus and getting a vaccination costs money, so:
  - If you get the vaccine and are infected, you will have spent some money but also survived (500).
  - If you get the vaccine but are not infected by the virus, you will have wasted money (-10).
  - If you do not get the vaccine and are infected by the virus, you will die (-5000).
  - If you do not get the vaccine, but you are not infected by the virus, you'll have saved some money (10).
- But now suppose that the probability of being infected after being exposed to the virus is $\frac{1}{3}$.
- In that case, *do what is likely to work best* recommends that you do not get the vaccine!

Chapter 7

Probability Theory

7.1 Probability and Measures

- A probability function is a **normalized measure** over a **possibility space**. So, what does that mean?

Measures

- A measure is a function from 'regions' of some space to non-negative numbers — which has the following property:
  - If A is a region that divides into regions B_1 ... B_n, then the *measure* of A is the sum of the measures of B_1 ... B_n.
- An example of a measure function m:
  - Let R be a set of Real Madrid players and let G be a set of numbers representing goals scored. Consider a mapping from e.g. the power of R into G representing a mapping from sets of players to the number of goals scored (by those players). This would be an example of a **measure**.

$$m\colon \mathcal{P}(R) \longmapsto G$$

  - So, in order for a function to be a measure, it must be **additive**. Additivity is defined as follows:

If $A \cap B = \varnothing$ (if A and B are disjoint), then $m(A \cup B) = m(A) + m(B)$.

  - So, if m is additive, it will satisfy the following:

Also called *countable additivity* or *σ-additivity*.

For any disjoint sets $A_1, A_2 \ldots A_n$,

$$m\Big(\bigcup_{n=1}^{N} A_n\Big) = \sum_{n=1}^{N} m\big(A_n\big)$$

- Let’s consider an example:

REAL MADRID PLAYER	GOALS SCORED
Gonzalo Higuaín	133
Ángel Di María	46
Cristiano Ronaldo	272
Karim Benzema	128
Mesut Özil	50

  - Since Higuaín and Di María are the only Argentinian players (in the list above), suppose we apply m to the set A of Argentinian players where $A \subseteq R$.
  - The value of $m(A)$ should then be 179 since this is the combined number of goals scored by Higuaín and Di María.
- In short, a **measure** is always additive over subregions (of the domain).

Normalized Measures

- A **normalized** measure is a function m where the value that m assigns to the entire domain is 1.
- We can *normalize* any measure by simply dividing each value in the range by the value of the universe, i.e. in this case dividing number of goals scored n (for each individual i) by the total number of goals scored (629).

REAL MADRID PLAYER	GOALS SCORED
Gonzalo Higuaín	$\frac{22}{100} = \frac{11}{50}$
Ángel Di María	$\frac{7}{100}$
Cristiano Ronaldo	$\frac{43}{100}$
Karim Benzema	$\frac{20}{100} = \frac{5}{100}$
Mesut Özil	$\frac{8}{100} = \frac{2}{25}$

Formal Definition

- Here is a slightly simplified but formal definition of a measure.

If a domain D satisfies 1. and 2., D is a so-called σ-algebra.

  - A measure is a function m that satisfies the following conditions:
    1. The domain D is a set of sets.
    2. The domain is closed under the following set theoretic operations: union, intersection, and complementation (with respect to the relevant universe U). So, if $A \in D$ and $B \in D$, then $A \cup B \in D$, $A \cap B \in D$, and $U \backslash A \in D$.
    3. The range is a set of non-negative real numbers.
    4. The function is additive.

- From this, several important results can now be proved, cf. Weatherson (2011, 18) — the most important of which is the following

$$m(A) + m(B) - m(A \cap B) = m(A \cup B)$$

- This holds whether or not $A \cup B$ is empty.

Possibility Spaces

- The notion of a possibility space should already be fairly familiar, but to recap we will just think of possibility spaces as ways the world could be.
- We can represent possibility spaces in terms of propositions and their associated truth values
  - For example if p and q are the only propositions under consideration, there are four possibilities (or worlds if you will):

	p	q
w_1	T	T
w_2	T	F
w_3	F	T
w_4	F	F

- These possibilities form the foundation of the possibility space that is used to build a probability function.
- Since a probability function is a normalized measure, each of the possibilities must have probabilities, i.e. real numbers, that sum to 1 — for example:

	p	q	PROBABILITY
w_1	T	T	0.33
w_2	T	F	0.17
w_3	F	T	0.42
w_4	F	F	0.08

- Since this is a function that maps each of the w's above to an associated real number and the sum of all these numbers is 1, we have a normalized measure over a possibility space — viz. a probability function.
  - **Note**. The probability of individual propositions can now be calculated by suming all the possibilities where that proposition is true.

    — Hence the probability of p, $\Pr(p)$, is $0.33 + 0.17 =$ **0.50**

    — And the probability of q, $\Pr(q)$, is $0.33 + 0.42 =$ **0.75**

7.2 Propositions and Probabilities

- While it is useful to think of probabilities in terms of truth tables, we have to be a bit careful.
- We generally assume that when we are considering n propositions, there are 2^n possibilities, but this is not always the case — consider the propositions below.

  A = Alfred is taller than Betty

  B = Betty is taller than Carlos.

  C = Carlos is taller than Alfred.

	A	B	C
1.	T	T	T
2.	T	T	F
3.	T	F	T
4.	T	F	F
5.	F	T	T
6.	F	T	F
7.	F	F	T
8.	F	F	F

- The problem here is that line 1. is not a genuine possibility!
- Keeping a careful eye on such cases, we define the notions below as follows.

  - **Logical Equivalence**
    Two sentence A and B are logically equivalent if and only if they have the same truth value at every line in a truth table *that represents a real possibility* — i.e.

	A	$\neg\neg A$
1.	T	T
2.	F	F

  - **Entailment**
    Sentences A_1, A_2 ... A_n entail a sentence B if and only if at every line which (a) represents a real possibility and (b) where A_1, A_2 ... A_n is true, B is also true. For example, $(A \wedge B)$ entails A:

	A	B	$(A \wedge B)$
1.	T	T	T
2.	T	F	F
3.	F	T	F
4.	F	F	F

- **Logically Disjoint Sentences**
  Two sentences A and B are logically disjoint if and only if there is no line in the truth table which (a) represents a real possibility and (b) is a line where both A and B are true.

	A	$\neg A$
1.	T	F
2.	F	T

- If some line in a truth table does not represent a real possibility, it is assigned probability 0.
- A truth table containing lines A_1, A_2 ... A_n represents a normalized measure only if the probability of $A_1 + A_2 + ... + A_n$ sum to 1, hence some adjusting might be necessary in certain cases.

We'll see in a minute that one of the basic axioms of probability theory is that logical truths have probability 1.

7.3 Axioms of Probability Theory

- We now have the notion of a normalized measure and we have defined probability functions in terms of these. We now turn to probability theory more generally—in particular, the probability of complex sentences.
- A probability function is a normalized measure function that takes sentences as inputs, has numbers in the [0,1] interval as output, but also satisfies the following constraints:
  1. If A is a logical truth, then $\Pr(A) = 1$
  2. If A and B are logically equivalent, then $\Pr(A) = \Pr(B)$.
  3. If A and B are logically disjoint (i.e. if $\neg(A \wedge B)$ is a logical truth) then $\Pr(A \vee B) = \Pr(A) + \Pr(B)$.
- From these axioms, several important results can now be proved.

P1. $\Pr(A) + \Pr(\neg A) = 1$

Proof.

- $(A \vee \neg A)$ is a logical truth.
- So, from Axiom 1 it follows that $\Pr(A \vee \neg A) = 1$.
- A and $\neg A$ are logically disjoint, so from Axiom 3 it follows that:

$$\Pr(A \vee \neg A) = \Pr(A) + \Pr(\neg A)$$

- Hence, $\Pr(A) + \Pr(\neg A) = 1$.

P2. If A is a logical falsehood, $\Pr(A) = 0$.

Proof.

- If A is a logical falsehood, $\neg A$ is a logical truth.

- So, by Axiom 1, $\Pr(\neg A) = 1$. Since $\Pr(A) + \Pr(\neg A) = 1$, it follows that:

$$\Pr(A) = 1–\Pr(\neg A)$$

- Thus, $\Pr(A) = 0$.

P3. $\Pr(A) + \Pr(B) = \Pr(A \vee B) + \Pr(A \wedge \mathrm{B})$

Proof. (in three steps)

1. Note that A is logically equivalent to $(A \wedge B) \vee (A \wedge \neg B)$.

- By Axiom 2: $\Pr(A) = \Pr((A \wedge B) \vee (A \wedge \neg B))$.
- $(A \wedge B)$ and $(A \wedge \neg B)$ are disjoint, hence it follows that:

$$\Pr(A \wedge B) \vee (A \wedge \neg B) \;=\; \Pr(A \wedge B) + \Pr(A \wedge \neg B).$$

▸ Hence, $\Pr(A) = \Pr(A \wedge B) + \Pr(A \wedge \neg B)$.

2. $(A \vee B)$ is equivalent to $(B \vee (A \wedge \neg B))$.

- B and $(A \wedge \neg B)$ are disjoint.
- Hence, $\Pr(B \vee (A \wedge \neg B)) = \Pr(B) + \Pr(A \wedge \neg B)$.
- From this it follows that:

$$\Pr(A \vee B) = \Pr(B) + \Pr(A \wedge \neg B)$$

3. If we add $\Pr(A \wedge B)$ to both sides of the equation above we get:

$$\Pr(A \vee B) + \Pr(A \wedge B) = \Pr(B) + \Pr(A \wedge \neg B) + \Pr(A \wedge B)$$

- But, we have already proved that $\Pr(A \wedge \neg B) + \Pr(A \wedge B) = \Pr(A)$.
- Hence we can substitute as follows:

$$\Pr(A \vee B) + \Pr(A \wedge B) = \Pr(B) + \Pr(A)$$

- And this is equivalent to:

$$\Pr(A) + \Pr(B) = \Pr(A \vee B) + \Pr(A \wedge B)$$

7.4 Conditional Probability

- Sometimes we might be interested not just in the probability of some proposition A, but the probability of A **given** (or **conditional on**) some other proposition B.
- This is called conditional probability: The probability of A given B — or the probability of A conditional on B. We write such conditional probabilities as follows:

$$\Pr(A|B)$$

- An example:

> Suppose we're deciding to bet on whether Real Madrid will score a goal. However, if Ronaldo does not play, we know that Real Madrid are much less likely to score. So, we're only going to bet on the assumption that Ronaldo plays. As a result, we're interested in the probability that Real Madrid will score a goal conditional on Ronaldo playing.

- We can think of conditional probabilities in terms of the elimination of possibilities. So, suppose that:
  - A = Real Madrid scores a goal.
  - B = Ronaldo plays.
- And suppose we have the following probability distribution.

	A	B	PROBABILITY
w_1	T	T	0.80
w_2	T	F	0.10
w_3	F	T	0.05
w_4	F	F	0.05

- If we are looking to determine the conditional probability of A given B, viz. $\Pr(A|B)$, we start by assigning the possibilities where B is false 0.

	A	B	PROBABILITY
w_1	T	T	0.80
w_2		F	~~0.10~~ 0
w_3	F	T	0.05
w_4		F	~~0.05~~ 0

- We are now left with something which is not a normalized measure, since the remaining possibilities do not sum to 1. The remaining values only sum to the probability of B — so we need to *re-normalize*.
- We can re-normalize by dividing the value of the remaining possibilities by $\Pr(B)$, viz. 0.8 ÷ 0.85 and 0.05 ÷ 0.85.
- Hence, we get:

	A	B	PROBABILITY
w_1	T	T	0.94
w_2		F	0
w_3	F	T	0.06
w_4		F	0

- Generally, the probability of A conditional on B can be calculated as follows.

$$\Pr(A|B) = \frac{\Pr(A \wedge B)}{\Pr(B)}$$

The probability of $A \wedge B$ is $A \times B$ (when A and B are *independent* — independence is discussed later.)

Calculating Conditional Probabilities

- When we are interested in $\Pr(A|B)$, it is often easier to calculate $\Pr(B|A)$.
    - For example, suppose we are interested in knowing the probability of drawing the ace of diamonds from a deck of cards conditional on us having already drawn a red card. It is obviously a lot easier to figure out the probability of drawing a red card conditional on having drawn the ace of diamonds.
- However, using the famous theorem due to Thomas Bayes, it is in fact possible to calculate $\Pr(A|B)$ using $\Pr(B|A)$, namely as follows:

$$\textbf{Bayes Theorem} \qquad \Pr(A|B) = \frac{\Pr(B|A) \times \Pr(A)}{\Pr(B)}$$

- Notice that $\Pr(B) = \big(\Pr(B|A) \times \Pr(A)\big) + \big(\Pr(B|\neg A) \times \Pr(\neg A)\big)$, so the following is an equivalent formulation of Bayes Theorem.

Equivalent formulation of Bayes' Theorem

$$\textbf{Bayes Theorem} \qquad \Pr(A|B) = \frac{\Pr(B|A) \times \Pr(A)}{(\Pr(B|A) \times \Pr(A)) + (\Pr(B|\neg A) \times \Pr(\neg A))}$$

- Let's consider an example.
    - Take a fair die. Suppose we wanted to calculate the probability rolling a 3 conditional on the roll being odd, i.e. $\Pr(A|B)$.

        A = The number rolled is a 3.

        B = The number rolled is odd.
    - The probability of the number rolled being odd when the roll was a 3 is obviously 1, viz. $\Pr(B|A) = 1$
    - The probability of rolling a 3 is $\frac{1}{6}$, viz. $\Pr(A) = \frac{1}{6}$
    - The probability of the number rolled being odd is $\frac{1}{2}$, viz. $\Pr(B) = \frac{1}{2}$.
    - So, using Bayes Theorem:

$$\begin{aligned} \Pr(A|B) &= \frac{\Pr(B|A) \times \Pr(A)}{\Pr(B)} \\ &= \frac{1 \times \frac{1}{6}}{\frac{1}{2}} \\ &= \frac{\frac{1}{6}}{\frac{1}{2}} \\ &= \frac{1}{3} \end{aligned}$$

    - Voilá.
- This example demonstrates an important point: the fact that the conditional probability of some proposition B given A is high (even 1), does not in any way indicate that the conditional probability of A given B is equally high (or high at all).

Another Example: Test Reliability

- Suppose we are trying to determine the probability of some patient A having a disease D conditional on the test coming back positive.
  - 5% of patients have D.
  - If patient a has D, the test returns a positive result 80% of the time.
  - If a does not have D, the test returns a negative result 90% of the time.
- We now use Bayes Theorem to calculate the conditional probability. Let:

  A = Patient a has disease D.
  B = The test returns a positive result.

- This yields the following *prior* probabilities.
  - $\Pr(A) = 0.05$
  - $\Pr(\neg A) = 0.95$
  - $\Pr(B|A) = 0.8$
  - $\Pr(B|\neg A) = 0.1$
- The calculation now proceeds as follows:

$$\begin{aligned}
\Pr(A|B) &= \frac{\Pr(B|A)\times\Pr(A)}{\Pr(B|A)\times\Pr(A) + \Pr(B|\neg A)\times\Pr(\neg A)} \\
&= \frac{0.8\times0.05}{0.08\times0.05 + 0.1\times0.95} \\
&= \frac{0.04}{0.04 + 0.095} \\
&= \frac{0.04}{0.135} \\
&\approx 0.3
\end{aligned}$$

- In conclusion, the chances of having the disease conditional on having tested positive is actually less than $\frac{1}{3}$.

7.5 Conditionalization

- We can distinguish between two concepts relating to conditional probabilities.

(H1) The probability of hypothesis H given evidence E.

(H2) The new probability of hypothesis H after evidence E has come in.

- It is important to recognize that these are distinct concepts.

- (H1) is a *static* concept that simply states the probability of H given E. It does not say anything about whether E obtains.
- (H2) is a *dynamic* concept that states something about *what to do* when evidence E is acquired.

- We will use $\Pr(H|E)$ for the concept in (H1) and $Pr_E(H)$ for the concept in (H2).

Conditionalizing on the Evidence

- Across several disciplines, e.g. economics, statistics, and mathematics, it is generally assumed that *rational* agents (should) update beliefs by *conditionalization*.
- For example, when a rational agent acquires some evidence E for hypothesis H, the rational agent simply replaces $\Pr(H)$ with $\Pr(H|E)$.
  - **Example**
    If the probability of H before was 0.5, but the probability of H given E was 0.9, then when E comes in, the probability of H is adjusted to 0.9.
- Hence, it is assumed that: $\Pr(H|E) = Pr_E(H)$.
- **An Argument for the Connection Between (H1) and (H2)**.

A1. — Suppose you draw two cards from a deck of cards (without replacement). There are 52 cards and 13 hearts, so the odds of drawing a heart is: $\frac{13}{52} = \frac{1}{4}$

— Suppose you draw a heart on the first draw. There is now one card and one heart less in the deck, so the probability of drawing a second heart on the second draw is then $\frac{12}{51}$. Hence, the probability of drawing a heart conditional on already having drawn a heart is $\frac{12}{51}$.

— In other words, when trying to work out $\Pr(H|E)$, it seems natural to pretend that we are trying to work out $Pr_E(H)$ and then simply stop pretending when we are done.

— That is, it seems natural to simply assume that calculating $\Pr(H|E)$ just is to evaluate what probability that we would assign to H when E comes in.

7.6 Probabilities: Independence

- The probability of some proposition A may or may not depend on the probability of another proposition B. In informal terms, if the probability of A does not depend on the probability of B, we say that A and B are *independent*.
- However, independence is formally defined as follows:

While this definition looks *asymmetric*, we will show that it is in fact symmetric.

Propositions A and B are **independent** iff $\Pr(A|B) = \Pr(A)$.

- **Example**
  You draw a card from a standard deck of cards.

- · A = Card drawn is a face card. · B = Card drawn is a heart.

- This yields the following probabilities:

$$\begin{aligned}\Pr(A) &= \tfrac{3}{13}\\ \Pr(B) &= \tfrac{1}{4}\\ \Pr(B|A) &= \tfrac{3}{12} = \tfrac{1}{4}\end{aligned}$$

- A and B are probabilistically independent because the likelihood of drawing a face card from a deck of 52 cards is the same as the probability of drawing a face card on the assumption that you have drawn a heart. This is easily demonstrated using Bayes Theorem.

$$\begin{aligned}\mathbf{Pr}(\boldsymbol{A}|\boldsymbol{B}) &= \frac{\Pr(B|A)\times\Pr(A)}{\Pr(B)}\\ &= \frac{\frac{1}{4}\times\frac{3}{13}}{\frac{1}{4}}\\ &= \frac{\cancel{\frac{1}{4}}\times\frac{3}{13}}{\cancel{\frac{1}{4}}}\\ &= \frac{3}{13} \qquad = \mathbf{Pr}(\boldsymbol{A})\end{aligned}$$

- The definition of independence is symmetric which can be proved as follows:

  - **Proof**: $\Pr(A|B) = \Pr(A) \longleftrightarrow \Pr(B|A) = \Pr(B)$

$$\begin{aligned}\mathbf{Pr}(\boldsymbol{A}|\boldsymbol{B}) = \mathbf{Pr}(\boldsymbol{A}) &\leftrightarrow \frac{\Pr(A\wedge B)}{\Pr(B)} = \Pr(A)\\ &\leftrightarrow \Pr(A\wedge B) = \Pr(A)\times\Pr(B)\\ &\leftrightarrow \frac{\Pr(A\wedge B)}{\Pr(A)} = \Pr(B)\\ &\leftrightarrow \mathbf{Pr}(\boldsymbol{B}|\boldsymbol{A}) = \mathbf{Pr}(\boldsymbol{B})\end{aligned}$$

7.7 Correlation vs. Causation

- It is very important to distinguish between **correlations** (i.e. probabilistic dependence) and **causal dependence**.

  - · **Correlations/Probabilistic Dependence**
    Above, we proved the following:

$$\Pr(A|B) = \Pr(A) \quad\longleftrightarrow\quad \Pr(B|A) = \Pr(B)$$

Hence, probabilistic dependence is a *symmetric* notion.

- For example, if A is (probabilistically) independent from B then B is (probabilistically) independent from A.
- Or in other words, if A is not correlated with B, then B is not correlated with A.

- **Causal Dependence**
  If A causes B, B is causally dependent on A.
- This is an *asymmetric* notion of dependence: That A causes B does not entail that B also causes A (in fact, it seems to entail the negation).

Correlations withouth Causal Dependence

- Many things are correlated, but not causally dependent — here are a couple of examples:

  1. Having a runny nose is correlated with having a sore throat, but a runny nose doesn't *cause* a sore throat.
  2. Having cancer is correlated with being bald, but cancer doesn't *cause* baldness.
  3. That the train arrives at the station is correlated with the station clock showing 12.00. But the clock's showing 12.00 doesn't *cause* the train to arrive.
  4. Etc.

- When two propositions (or events really) A and B are correlated (viz. probabilistically dependent), it will not be the case that $\Pr(A|B) = \Pr(A)$.
- To see why, let's work through an example.

  - Suppose that Jack and Bill are fans of Real Madrid and happy whenever Real Madrid wins.
  - So, when Jack is happy, this is some evidence that Real Madrid has won, and hence some evidence that Bill is happy too.
  - In other words, there is a probabilistic connection between Jack's being happy and Bill's being happy.
  - But Jack's being happy is not *the cause* of Bill's being happy (let's at least suppose) — so there is no causal connection between the two.

- Let's use some formalism to work out the probabilistic dependence:

  A = Real Madrid wins.

  B = Jack is happy.

  C = Bill is happy

  - Suppose that Real Madrid have 0.6 chance of winning.
  - If Real Madrid wins, the probability that Jack and Bill are happy is 1.

- If Real Madrid loses, the probability that Jack and Bill are happy is 0.5, but conditional on Real Madrid losing, Jack's being happy and Bill's being happy are probabilistically independent.

- In other words, if Real Madrid loses, Jack is happy 50% of the time and Bill is happy 50% of the time.

	A	B	C	Pr
1.	T	T	T	0.6
2.	T	T	F	0
3.	T	F	T	0
4.	T	F	F	0
5.	F	T	T	0.1
6.	F	T	F	0.1
7.	F	F	T	0.1
8.	F	F	F	0.1

- This now yields the following results:

$$\begin{array}{rcccc|l} \Pr(B) & = & 0.6 + 0.1 + 0.1 & = & 0.8 & \text{(sum of rows 1,2,5,6)} \\ \Pr(C) & = & 0.6 + 0.1 + 0.1 & = & 0.8 & \text{(sum of rows 1,3,5,7)} \\ \Pr(B \wedge C) & = & 0.6 + 0.1 & = & 0.7 & \text{(sum of rows 1,5)} \end{array}$$

- Remember, the probability of B conditional on C equals the probability of the conjunction of B and C divided by B:

$$\begin{array}{rcl} \Pr(B|C) & = & \frac{\Pr(B \wedge C)}{\Pr(C)} \\ & = & \frac{0.7}{0.8} \\ & = & 0.875 \end{array}$$

- Since $\Pr(B) = 0.8$ and $\Pr(B|C) = 0.875$, it follows that the probability of B conditional on C is greater than the probability of B.

$$\Pr(B|C) > \Pr(B)$$

- This shows that B and C are probabilistically dependent—viz. B and C are correlated, because conditionalizing on C raises the probability of B.

- This can be explained as follows: If C is true, this raises the probability of one of the probable causes of C and that cause is also a possible cause of B. Hence, the probability of B is increased.

Common Causes

- When two propositions (or events really) A and B are correlated, this arguably implies that either (i.), (ii.), or (iii.) holds.

  i. A causes B.

  ii. B causes A.

  iii. A and B have a common cause.

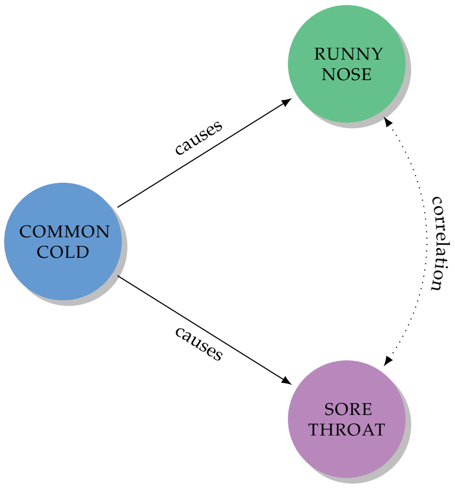
- Controlling for common causes can be a useful way of determining whether a correlation is causal.
- To illustrate, consider again the example of runny noses and sore throats.

  $B = a$ has a runny nose.

  $C = a$ has a sore throat.

- B and C are correlated but appear to have a common cause, namely A.

  $A = a$ has a cold.



- Since a cold is a probable common cause, we should consider what happens to the correlation when we "screen off" cases where the subject has a cold from cases where the subject does not.
- Doing this, we find that the correlation "disappears" in the following sense:
  - In cases where a has a cold, that a has a runny nose does not make it *any more* likely that a has a sore throat.
  - In cases where a does not have a cold, that a has a runny nose does not make it *any more* likely that a has a sore throat.

- So, while (1) initially holds:

(1) $\Pr(B|C) > \Pr(B)$

- Controlling for a common cause reveals that (2) and (3) hold too:

(2) $\Pr(B|(C \wedge A)) = \Pr(B|A)$

(3) $\Pr(B|(C \wedge \neg A)) = \Pr(B|\neg A)$

- In other words, this reveals that B and C are not causally dependent, because when A is controlled for, C has no impact on the probability of B and vice versa.
- This kind of probabilistic data is useful to rule out so-called *spurious* correlations. For example, suppose someone made the following argument:

> **Teenagers, Behavior, and Video Games**
> Teenagers who frequently play video games tend to be antisocial. So, video games cause antisocial behavior.

- While there might be a correlation between antisocial behavior and playing violent video games, this is quite likely a *spurious* correlation.
- If other relevant factors were controlled for, it seems likely that we would find that there is no causal connection between the two.

Chapter 8

Utility and Probability

8.1 Utilities and Expected Values

- A random variable, X, is a variable that takes different values with respect to different possibilities.
- For example, suppose 14 people have been asked to predict the outcome of the Wimbledon final.
  - 9 people predict that Federer will win.
  - 5 people predict that Federer will lose.
- In this case, X takes the following values:

$$X = \begin{cases} 9 & \text{if Federer wins.} \\ 5 & \text{if Federer loses.} \end{cases}$$

- Given a random variable X and a probability distribution, we can calculate **the expected value** of X.
- The expected value of X, which we write as $Exp(X)$, is the sum of the numbers obtained by multiplying each value of X with the probability of that state obtaining.
  - A = Federer wins
  - $\Pr(A) = 0.8$

	A	PROBABILITY	PREDICTIONS (P)	PR(A) × P
1.	T	0.8	9	7.2
2.	F	0.2	5	1

Exp(X)
8.1
(7.1 + 1)

- The expected value here represents a general expectation of how many people would make correct predictions on average given these probabilities.

8.2 Maximise Expected Utility

- The orthodox view in decision theory is that one should **always maximise expected utilities**. Let's refer to this principle as the MAXIMISE EXPECTED UTILITIES rule.
- The approach to decision making which takes correct decisions to be calculated in terms of expected utilities is generally called *expected utility theory*.
- In expected utility theory, it is standardly assumed that an agent is rational only if the agent makes decisions which conforms to the MAXIMISE EXPECTED UTILITIES rule.
- Consider again the case involving **Airline 1** and **Airline 2**
  - **Airline 1** is cheap but very unreliable in bad weather. When the weather is bad, the luggage tends to be delayed.
  - **Airline 2** is expensive but reasonably reliable in bad weather. Rarely when the weather is bad is the luggage delayed.

· A = Weather is good.	· $\Pr(A) = 0.8$
· B = Weather is bad.	· $\Pr(B) = 0.2$

- Now consider the following decision table (cardinal utilities).

CASE I	*A*	*B*
Airline 1	20	4
Airline 2	6	5

- Calculating the expected values in CASE I for *A* and *B* respectively:

$$\begin{aligned} \mathit{Exp}(\textbf{Airline 1}) &= U(\text{Airline } 1|A) \times \Pr(A) \; + \; U(\text{Airline } 1|B) \times \Pr(B) \\ &= 0.8{\times}20 \; + \; 0.2{\times}4 \\ &= 16 + 0.8 \\ &= \mathbf{16.8} \end{aligned}$$

$$\begin{aligned} \mathit{Exp}(\textbf{Airline 2}) &= U(\text{Airline } 2|A) \times \Pr(A) \; + \; U(\text{Airline } 2|B) \times \Pr(B) \\ &= 0.8{\times}6 \; + \; 0.2{\times}5 \\ &= 4.8 + 1 \\ &= \mathbf{5.8} \end{aligned}$$

- So, the MAXIMISE EXPECTED UTILITY rule recommends using **Airline 1** as this has the highest expected utility.
- As a contrast, let's consider the revised version of the case that we discussed in the lecture 7.

- Assume that the probability of each state is the same.
  - **Airline 1** is cheap but very unreliable in bad weather. When the weather is bad, planes tend to crash.
  - **Airline 2** is expensive but reasonably reliable in bad weather. Sometimes when the weather is bad is the luggage delayed.
- This now yields the following decision table.

CASE II	*A*	*B*
Airline 1	20	-1000
Airline 2	6	5

- Calculating the expected values in CASE II for *A* and *B* respectively:

$$\begin{aligned} \mathit{Exp}(\textbf{Airline 1}) &= U(\text{Airline } 1|A) \times \Pr(A) \ + \ U(\text{Airline } 1|B) \times \Pr(B) \\ &= 0.8{\times}20 \ + \ 0.2{\times}{-1000} \\ &= 16 - 200 \\ &= \mathbf{-184} \end{aligned}$$

$$\begin{aligned} \mathit{Exp}(\textbf{Airline 2}) &= U(\text{Airline } 2|A) \times \Pr(A) \ + \ U(\text{Airline } 2|B) \times \Pr(B) \\ &= 0.8{\times}6 \ + \ 0.2{\times}5 \\ &= 4.8 + 1 \\ &= \mathbf{5.8} \end{aligned}$$

- Now the MAXIMISE EXPECTED UTILITY rule recommends using **Airline 2** as this has the highest expected utility.
- Of course, different results could be obtained by either (a) changing the utilities associated with each outcome and (b) changing the probability distributions.
- However, notice that different results can be obtained even without changing the ordinal ranking of the outcomes, i.e.

Airline 1		**Airline 2**		**Airline 2**		**Airline 1**
Good Weather	>	Good Weather	>	Bad Weather	>	Bad Weather

- For example, just make bad weather sufficiently improbable (for example, just suppose that it would have to be particularly bad weather for it to have an effect). Doing that, we could make the MAXIMISE EXPECTED UTILITY rule recommend using **Airline 1** in CASE II.

Homework Exercise: Calculate what probability bad weather would have to have in **Case II** in order for the maximise expected utility rule to recommend **Airline 1**.

8.3 Properties of the Maximise Expected Utility Rule

- When using the MAXIMISE EXPECTED UTILITY rule, the option picked is simply the option with the highest number (where this number is the calculated expected utility).

- This rule has a number of interesting properties, namely:
  - Guaranteed to be transitive.
  - Satisfies the independence of irrelevant alternatives
  - Never recommends choosing dominated options.
- Let's write the expected utility of a choice A as $Exp(U(A))$.
  - **Transitivity**
    If A is recommended over B, then $Exp(U(A)) > Exp(U(A))$, and if B is chosen over C, then $Exp(U(B)) > Exp(U(A))$.

    Since the output of $Exp(U(\cdot))$ is a number and '>' defined over numbers is transitive, it follows that $Exp(U(A)) > Exp(U(C))$.
  - **Independence of Irrelevant Alternatives**
    If A is chosen over B and C, then $Exp(U(A)) > Exp(U(B))$ and $Exp(U(A)) > Exp(U(C))$. Given this, then regardless of whether the choices are between A and B only, or between A and C only, A will still be chosen.

    Notice that numbers are totally ordered, i.e. either $x > y$, $y > x$, or $x = y$. Since choices are associated with numbers, a similar relation holds between choices.
  - **No Dominated Choices**
    Assume that A dominates B. Let $U(A|S_i)$ denote the utility of A in state S_i. If A dominates B, this means that for all i, $U(A|S_i) \geq U(B|S_i)$.

    Now, remember, the expected utility of A and B are calculated as follows:

1. $Exp(U(A)) = \Pr(S_1)\times U(A|S_1) + \Pr(S_2)\times U(A|S_2) + ... + \Pr(S_n)\times U(A|S_n)$
2. $Exp(U(B)) = \Pr(S_1)\times U(B|S_1) + \Pr(S_2)\times U(B|S_2) + ... + \Pr(S_n)\times U(B|S_n)$

Assuming that $\Pr(S_i) > 0$ for at least one i, then from dominance it follows that each term in row 1. is at least as great as the corresponding term below it — and for at least one term in row 1., it is greater than the corresponding term below it. Since $Exp(U(\cdot))$ is always the sum of n terms, it follows that $Exp(U(A)) > Exp(U(B))$. Hence, if A dominates B, A has a higher expected utility.

8.4 A More General Version of Dominance

- Here is what would appear to be a reasonable generalization of the dominance principle.
  - Assume our initial states are S where $S = \{S_1, S_2, ..., S_n\}$
  - Let B be a pair of sets of states, T_1 and T_2.
  - Assume that for all members S_i in S: $S_i \in T_1$ or $S_i \in T_2$ (but not in both).
- The generalized version of dominance says the following:

Generalized Dominance
If A is better than B in T_1 and A is better than B in T_2, then A is better than B in S.

- While this seems like a plausible principle, interestingly it is violated by some principles, e.g. the maxiaverage principle considered by (Weatherson, 2011, 52-53).
- The MAXIMISE EXPECTED UTILITY rule respects this generalized version of dominance, which seems good.

8.5 The Sure Thing Principle and the Allais Paradox

- Another principle that the MAXIMISE EXPECTED UTILITY rule satisfies is the so-called 'sure thing principle'.

Sure Thing Principle: If $AE \geq BE$ and $A\neg E \geq B\neg E$, then $A \geq B$.

- Some terminology to unpack this principle:
  - Read '$A \geq E$' as 'A is at least as good as B'.
  - 'AE' means 'A and E' — so '$AE \geq BE$' can be understood as a *preference* for A over B conditional on E, viz. a conditional preference.
  - So, the sure thing principle says that if A is at least as good as B conditional on E and conditional on $\neg E$, then A simply is at least as good as B.
- One worry about this principle (and hence a worry for the MAXIMISE EXPECTED UTILITY rule is a paradox formulated by Nobel Prize recipient Maurice Allais (1953).
- Suppose you were offered a gamble: a choice between A and B, but the result of your choice will depend on drawing a ball from an urn. The urn contains the following distribution of balls:

– 10 **Red** balls.	$\Pr(\textit{Red}) = {}^{1}\!/_{10}$
– 1 **Green** ball.	$\Pr(\textit{Green}) = {}^{1}\!/_{100}$
– 89 **Black** balls.	$\Pr(\textit{Black}) = {}^{89}\!/_{100}$

	Red	**Green**	**Black**
A	\$1,000,000	\$1,000,000	\$0
B	\$5,000,000	\$0	\$0

- In other words, this yields the following options:
  - **Choose A**: 11% chance of winning \$1,000,000.
  - **Choose B**: 10% chance of winning \$5,000,000.
- Needless to say, most people choose B over A here.
- Compare this to the following gamble: again a choice between A and B and the same urn:

	Red	Green	Black
C	$1,000,000	$1,000,000	$1,000,000
D	$5,000,000	$0	$1,000,000

- In other words, this yields the following options:
  - **Choose C**: 100% chance of winning $1,000,000.
  - **Choose D**: 10% chance of winning $5,000,000 and 89% chance of winning $1,000,000.
- Many people choose the sure thing here, namely *A*.
- The problem here is that one cannot consistently hold the following three views.
  - $B > A$
  - $C > D$
  - The Sure Thing principle holds.
- Remember, the Sure Thing principle says: If $AE \geq BE$ and $A\neg E \geq B\neg E$, then $A \geq B$.

  E = green or red ball is drawn. $\neg E$ = black ball is drawn.
- Now consider the following:
  1. $A\neg E$ is equivalent to $B\neg E$, as both yields an outcome of $0.
  2. Hence, if $AE > BE$, then by the Sure Thing principle, $A > B$.
  3. Equivalently, if $BE > AE$, then by the Sure Thing principle, $B > A$.
  4. Since most people think $B > A$, we conclude $BE > AE$.
  5. $C\neg E$ is equivalent to $D\neg E$, as both yields an outcome of $1,000,000.
  6. Hence, if $CE > DE$, then by the Sure Thing principle, $C > D$.
  7. Equivalently, if $DE > CE$, then by the Sure Thing principle, $D > C$.
  8. Since most people think $C > D$, we conclude $CE > DE$.
  9. But now notice that:
     — $AE = CE$
     — $BE = DE$
  10. That is, conditional on E, choosing between A and B is the same as choosing between C and D.
  11. Hence, it is inconsistent to hold that $BE > AE$ and $CE > DE$
- Another way of looking at the problem here is that the *expected amount of money* is higher in both *B* and *D*. So, the MAXIMIZE EXPECTED UTILITY rule (if we are liberally interpreting *utility* directly in terms of money) predicts choosing *B* and *D*.
- And so, if being a rational agent is defined in terms of conformity with the MAXIMISE EXPECTED UTILITY rule, then given that most people prefer *B* and *C*, expected utility theory deems most people irrational.

8.6 Interpretations of Probability

8.6.1 Probabilities as Frequencies

- The historical view of probability has been to identify probabilities with frequencies.
  - For example, if the aim is to determine the probability of student *b* catching influenza, looking at the frequency with which students catch influenza seems like a good start.
  - If the proportion of students that catch influenza each winter is, say, $^1/_{10}$ — we might conclude that the probability of *a* catching influenza is also $^1/_{10}$.
- The immediate problem with interpreting probabilities in terms of frequencies is that when additional information is added, probabilities change. For example, suppose you're informed about one or more of the following facts.
  - · *b* has not had an influenza shot.
  - · *b* is working in a hospital.
  - · *b* is particularly susceptible to influenza.
- Intuitively, given these facts, one should not maintain that the probability that *b* will catch influenza is a mere $^1/_{10}$.
- This problem is an instance of two more general problems.
  - · **The Reference Class Problem**
    If we are considering the probability of *b* being *F*, then the references classes that *b* belongs to might differ with respect to the frequency of the members of that class having the property *F*.
  - · For example,
    - — Class of foreign students that catch influenza each winter: $^1/_5$
    - — Class of first year students that catch influenza each winter: $^1/_{10}$
    - — Class of students from high income families that catch influenza each winter: $^1/_{20}$
  - · This problem might be solved by simply looking at the most narrowly defined reference class. However, this leads to:
  - · **The Single Case Problem**
    Often, we are interested in the probability of a one off event, e.g. the probability that Chris Christie will be the Republican nominee in 2016, that Goldman Sachs will default etc.
  - · However, for such one off events, the relevant frequencies are simply 1 or 0, which is not helpful.

8.6.2 Degrees of Beliefs — Bayesianism

- The term 'Bayesianism' is a general term for a collection of positions in various related fields which centers on the interpretation of probability as something like *degrees of belief* or *credences*.
- Bayesians typically agree about the following three things:

  1. There is an important mental attitude of *degree of belief* that can be characterized in terms of numerical values in the [0,1] interval.
  2. If an agent is perfectly rational, her degrees of belief obeys the axioms of probability theory.
  3. Conditionalization is the standard way that beliefs change over time.

- On a Bayesian interpretation of probability, the probability of some proposition P is simply how confident an agent is that P obtains.
- This of course raises the question of which numerical values an agent should assign to her beliefs, and why.
- The most famous type of argument for Bayesianism is the so-called *Dutch Book* argument.

  Fair Bets

  For now, we assume that utility can be interpreted directly in terms of monetary value.

  - Imagine you are offered a bet that pays \$1 if P is true. How much should you be willing to pay for this bet?
  - Where A is the action of taking the bet and $U(A)$ is the utility of that action, the bet would have the payout structure below:

  $$U(A)\begin{cases} 1-\Pr(P) & \text{if } P \\ -\Pr(P) & \text{if } \neg P \end{cases}$$

  - The expected utility of A is:

  $$\begin{aligned} Exp(U(A)) &= \Pr(P)\times U(A|P) \ + \ \Pr(\neg P)\times U(A|\neg P) \\ &= \Pr(P)\times(1-\Pr(P)) \ + \ \Pr(\neg P)\times(-\Pr(P)) \\ &= \Pr(P)\times(1-\Pr(P)) \ + \ (1-\Pr(P))\times(-\Pr(P)) \\ &= \Pr(P)\times(1-\Pr(P)) \ - \ (1-\Pr(P))\times\Pr(P) \\ &= 0 \end{aligned}$$

  - So, if you pay \$Pr(P) for the bet, your expected return is 0.
  - If you pay more than \$Pr(P), your expected return is negative — and if you pay less than \$Pr(P), your expected return is positive.

- In sum, we measure an agent's degree of belief in P by considering the maximum price that the agent would pay for a bet that returns \$1 if P is true.
- There are several worrying features of this analysis.

Agent's Can Do No Wrong

- When it comes to credences, it seems that an agent is always right.
  - — If the agent takes the bet, this shows that the agent's degree of belief in P must be ≥ 0.5, and hence the agent maximises expected utility.
  - — If the agent declines the bet, this shows that the agent's degree of belief in P must be <0.5, and hence the agent maximises expected utility.
  - — So, in conclusion, the agent never makes a mistake.
- **Response**
  While no restrictions are placed on an agent's credence in a single proposition, there are some general constraints.
- It is generally assumed that an agent is rational only if she is not subject to a *Dutch book*.
- Avoiding *Dutch books* is a kind of consistency requirement. For example,
  - — Suppose some agent a is willing to accept a bet that pays \$1 if P for the price of \$0.60.
  - — But now suppose that a is also willing to accept a bet which pays \$1 if $\neg P$ for the price of \$0.60.
  - — In such a case, a is susceptible to a *Dutch book* — that is, a series of bets which guarantee that A loses money.
- That's bad, so if A is rational, she should avoid this.
- One way to avoid this is to ensure that one's credences respect the axioms of probability theory.
- I.e. if A's degree of belief in P is 0.6 (if \$0.6 is the maximum price she's willing to pay for a bet on P), then A's degree of belief in $\neg A$ should be 0.4.
- **Extending the Dutch book argument to Conditional Probabilities**
  - — Define a bet on A conditional on B as a bet that:
    - · Pays \$1 if A and B are true.
    - · Pays \$0 if A is false but B is true.
    - · The cost of the bet is returned if B is false.
  - — If we assume that $\Pr(A|B)$ specifies a fair price for this bet, the agent is subject to a Dutch book if she violates the following:

$$\Pr(A|B) = \frac{A \wedge B}{B}$$

- Dutch book arguments are also sometimes used to argue in favor of update by conditionalization — that is, it can be shown (omitting certain assumptions) that unless an agent updates by conditionalization, the agent is subject to a Dutch book.
- The other most prominent argument in favor of Bayesianism more generally are so-called Representation Theorems — however, we will leave those aside for now.

- **A Related Problem**
  A related worry about the Bayesian interpretation of probability as credences is that if an agent has manifestly crazy beliefs, viz. crazy credences, Bayesianism doesn't offer any resolution. For example,

  — Suppose that Frank has 0.9 credence in the proposition that Galatasaray will win the Champion's League.

  — Hence, Frank is willing to accept a bet that pays \$1 if Galatasaray wins the Champion's League at a ≤\$0.9 price.

- Given what Frank desires (viz. to win money) and given his credences, this is the correct thing to do for Frank — according to the Bayesian.
- Yet, intuitively, Frank should not accept such a bet, because Frank just has a manifestly crazy level of confidence in Galatasaray winning the Champion's League.

8.6.3 Evidential Probability

- The problem described in the previous section is, roughly, that the Bayesian interpretation of probability provides guidance with regards to what an agent should do **given the credences that the agent has**.
- But generally speaking, the more interesting question is what the agent should do — simpliciter.
- One suggestion to improve on the Bayesian picture is to continue to understand probability in terms of betting behavior, but add that it must be betting behavior **under perfect rationality**!
- Here, a perfectly rational agent takes into account the available evidence.
- So which credence an agent should have in Galatasaray winning the Champion's League depends on the maximum price that a perfectly rational agent would pay for a \$1 bet on Galatasaray winning.
- **Problem**

  — One immediate problem with appealing to perfect rationality is that it seems perfectly conceivable that rational people might evaluate available evidence in different ways.

  — I.e. some people might think that the evidence favors P and hence accept a 50/50 bet on P, whereas other people might find the evidence favors $\neg P$ and hence accept a 50/50 bet on $\neg P$.

  — Ideally, we should not be forced to conclude that one side in this debate is irrational.

- The notion of perfect rationality might not be the ideal way to go, but there is something to the idea of making evidence play a role in determining credences.
- While people might disagree about how strongly some evidence E supports a proposition P, one might think that there nevertheless is a fact about how strongly it supports P.

- Hence, we might say that what agent's should generally do is proportion their credences to the available evidence.
- Thus, the probability of a proposition P is determined by the maximum price that the agent would pay for a \$1 bet on P which in turn should be determined by the evidence available in favor of P.
- While there are still many problematic issues related to this way of analyzing probability — evidential probability — this is the notion that we rely on in the following.

8.6.4 Objective Chance

- A notion intuitively related to that of probability is the notion of *chance*.
- However, it is important to realize that chances are not evidential probabilities — here are some arguments for that claim.

  1. Events that took place in the past only have 1 or 0 chances. For example, if we considering whether some particle decayed in the past — it either did or it did not.

     However, many events in the past intuitively have an evidential probability i where $1 > i > 0$.

  2. Chances are objective. The evidential probability of P might differ from one agent to another, but this is not the case for objective chances. The objective chance of some particle decaying is independent of whatever evidence may be available.

Link Between Objective Chance and Evidential Probability

- There is intuitively a close link between objective chance and evidential probability.
- For example, if an expert physicist tells you the chance of some particle decaying in the next hour is 0.8, it seems that your credence in the proposition that the particle will decay in the next hour should be 0.8.
- This general idea is summed up in the principle below:

$Ch(P)$ = The objective chance that P

The Principal Principle $\Pr(P|Ch(P) = x) = x$

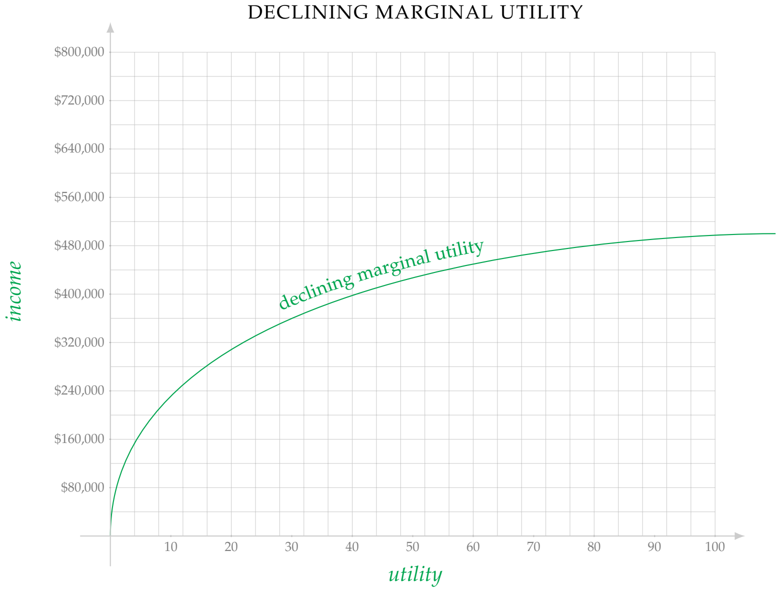
- The Principal Principle says: The evidential probability of P conditional on the chance of P being x is x.
- This seems like a reasonable principle, and it is hard to imagine counterexamples.
- It is important to notice that chances are not frequencies. For example, it may be possible to determine the chance of some particle not decaying in the next 10 years even if all particles of this kind have decayed in less than 5 years so far.
- In the following, we will often appeal to objective chance either implicitly or explicitly when setting up problems.

Chapter 9

More on Utility

9.1 Declining Marginal Utility

- In several of the decision problems considered so far, utilities have been directly equated with money. There are several reasons to be weary of conflating these — one being that the value of money depends how much money the agent already has.
- In more technical jargon, it is often said that money has **declining marginal utility**. In a graph, declining marginal utility would look something like this:



- The more money you earn, the less valuable each extra dollar you acquire is.

- One way to increase expected utility is to take out insurance. However, this is at the same time a way of decreasing one's monetary assets. Let's consider an example.

9.1.1 Insurance

- For the purposes of the following case, let's assume that the utility of x dollars is equal to $x^{\frac{1}{2}}$. Even if this is not a particularly plausible account of the relation between money and utility, it suffices to make a point.

  - Irene has an income of $57,600 pr. year.
  - There is a 5% chance that Irene will have her wages reduced to $16,900 pr. year.
  - Hence, Irene's expected yearly income is:

Expected Income

	57.600×0.95	+	16,900×0.05
=	54720	+	845
=	**55,565**		

  - Assuming that the utility Irene gets from an income of x dollars is $x^{\frac{1}{2}}$, the expected utility of Irene's income is:

Expected Utility

	$(57.600^{\frac{1}{2}})\times 0.95$	+	$(16{,}900^{\frac{1}{2}})\times 0.05$
=	228	+	6.5
=	**234.5**		

- Now suppose Irene is offered insurance against having her wages reduced. If she takes the insurance, she is guaranteed the income that she has now. The insurance costs $2,300.

  - Buying insurance guarantees Irene an income of $57,600 – $2,300 = $55,300.
  - This yields an expected utility of: $(55{,}300)^{\frac{1}{2}} \times 1 =$ **235.2**
  - In sum, the expected utility is higher than it was before she bought insurance.
  - Meanwhile, this is not a bad deal for the insurance company either:
    - Revenue: $2,300
    - 5% chance of $40,700 loss ($57,600–$16,900).
    - Expected monetary value: $2300 + (–$40.700 × 0.05) = $265
  - So, offering insurance is profitable for the insurance company.

9.2 Utility and Welfare

- The notion of utility is naturally associated with welfare, but there are several different theories about how to explicate that notion. There are e.g.

- Experience Based Theories
- Objective List Theories
- Preference Based Theories

9.2.1 Experience Based Theories of Welfare

- The most famous experience based theory of welfare is the theory associated with Jeremy Bentham, the father of *utilitarianism*. On this view, welfare is equal to having good experiences. I.e. if a person b has predominantly good experiences, b's welfare is high.
- Hence, on this view, the utility of an action is determined by how good of an experience it would lead to.

Problems

- There are several well known objections to experience based theories of welfare. Let's consider a couple.
  - **Problem I:** Nozick's Experience Machine
    Suppose a person b was wired up to a machine that produces the experiences of a good life. The machine is constructed such that it actually provides b with the **experience** of having loving and fulfilling relationships with friends and family, exploring the world, job success, kids etc.
  - Would b have had a good or bad life? Most people think not, since it would have been based entirely on an illusion.
  - **Problem II**: Varying Preferences
    Many people have different preferences when it comes to experiences. Some people enjoy rollercoaster rides and some people don't. Some people enjoy eating fois gras, and other people don't. What should the experience theorist say about this?
  - One option is to maintain that many people are simply systematically wrong about what good experiences are.
  - Another option is to argue that any experience always involves a first order experience and a second order experience. I.e. there is the experience of eating fois gras and then there is the experience of experiencing the eating of fois gras. Since these are different experiences, the thought is that the problem of varying preferences are strictly at the second order level.

9.2.2 Objective List Theories of Welfare

- A second approach to the analysis of welfare is so-called objective list theories.
- The general idea of objective list theories is, as the name suggests, that there is an objective list of things that generally make life better. Such a list might include:
  - Good health

- Loving relationships
- Adequate shelter, food, drink etc.
- Knowledge
- Engaging in rational activity
- Experiencing beauty
- Etc.

- Maximising either of these things will make your life better and hence whatever decisions you make should be decisions that lead to the maximisation of things on the list.

Problems

- Just like the experience based theories of welfare, objective list theories face several problems.

  - **Problem I**: Determining the List
    It is not obvious (a) what things should go on the list, (b) how that is supposed to be determined, (c) nor who is supposed to determine it. But without an answer to that question, the objective list theory is practically unusable with regards to decision making.
  - **Problem II**: Weighing the Strength of List Items
    Another problem is how to weigh the strength of various items on the list. In particular, with regards to decision theory, we need to be able to assign numerical values to various outcomes and this in turn requires an ordering of the items in terms of goodness.
  - It seems that the objective list theorist is forced to actually provide a complete ranking of the items on the list in order for the objective list to provide guidance with respect to action.

9.2.3 Preference Theories of Welfare

- Preference based theories of welfare start from the assumption that what is good for a person is a subjective matter — i.e. something that may vary from person to person.
- Hence, an outcome X is better than outcome Y if and only if the agent in question **prefers** X over Y.
- The preference theory of welfare avoids the problems discussed above concerning (a) varying preferences, and (b) which things to classify as welfare maximizing. It also deals with the problem of comparisons — whether some good x is better than y is simply relative to the preferences of various agents.

Problems

- However, there are also worries for the preference based theory of welfare.

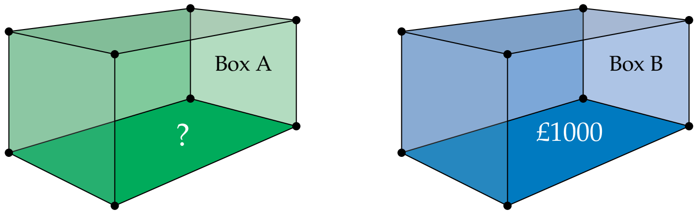
  - **Problem I**: Desiring Undesirable Things
    On the preference based theory of welfare, if an agent b prefers X over Y, this simply entails that X is better for b than Y.
  - But while a theory of welfare should not entail that people are systematically mistaken about what is good for them, there are undoubtedly cases where agents desire things that it is not in their best interest to desire. Such cases look like clear counterexamples to the preference based theory.
  - One possible modification of the preference based theory would be to maintain that an agent's welfare is maximised only by fulfilling the desires that the agent wished she had. But again, this might solve some of the problems, but counterexamples are rather easy to construct.
  - **Problem II:** Welfare of Groups
    Since welfare is a subjective notion on the preference based theory, evaluating the overall welfare of outcomes with respect to groups is in certain cases (though not all) problematic.
  - If some members of a group prefer X over Y and no member prefers Y over X, then X is said to be a **Pareto** improvement over Y. Also, if an outcome X is such that no other outcome Y is a Pareto improvement over X, then X is said to be **Pareto optimal**. Pareto improvements and Pareto optimality is, it would seem, straightforward to capture on the preference based theory.
  - Nevertheless, imagining cases where some members of a group prefer X over Y, but some other members prefer Y over X is easy — and in such cases, it is unclear what a theory of welfare based on subjective preferences is supposed to predict.
  - This is a problem, because there is a multitude of cases where, with respect to the welfare of a group, it intuitively seems like one outcome X is genuinely better than Y, but where there are differing preferences among the group members. It is difficult to see how this is supposed to be captured on a purely subjective and preferential understanding of welfare.

Chapter 10

Newcomb's Problem

10.1 Solutions to Newcomb's Problem

- We already introduced Newcomb's problem above, but here is a reminder.
- In front of you are two boxes **A** and **B**. You can see that **B** contains £1000, but you cannot see what is in **A**.



- You have two options:

  O1 Take **A**.

  O2 Take both **A** and **B** including the £1000.

- However, there is a catch (obviously).
  - A demon has predicted whether you will take just one box or two boxes. The demon is very good at predicting these things and has in the past made many predictions all of which have turned out to be correct.
  - If the demon predicts that you'll take both boxes, then she puts nothing in **A**. However, if the demon predicts that you'll take just one box, she'll put £1,000,000 in **A**.
- Hence, we have the following decision table.

	Predict 1 Box	*Predict 2 Boxes*
Take 1 Box	£1,000,000	£0
Take 2 Boxes	£1,001,000	£1,000

10.2 Two (potentially) Conflicting Decision Principles

- Robert Nozick has argued that Newcomb's problem introduces a challenge for two decision rules, namely:
  - Never choose dominated options.
  - Maximise expected utility.
- To appreciate why Newcomb's problem might pose a challenge, let's consider what action each of these decision rules recommend.
  - **Dominance**
    Since the 2-box option strongly dominates the 1-box option, respecting the prohibition against choosing dominated options requires that we take 2 boxes.
  - **Maximise Expected Utility**
    We have already proved that the MAXIMISE EXPECTED UTILITY rule never recommends choosing dominated options, so this rule will make the same recommendation.

    However, because Nozick relies on an alternative method for calculating expected utility, he predicts that the MAXIMISE EXPECTED UTILITY rule violates dominance. In particular, Nozick calculates the expected utility as follows:

1.	$Exp(U(A))$	$=$	$\Pr(S_1\|A)\times U(A\|S_1) + \Pr(S_2\|A)\times U(A\|S_2) + \ldots + \Pr(S_n\|A)\times U(A\|S_n)$
2.	$Exp(U(B))$	$=$	$\Pr(S_1\|B)\times U(B\|S_1) + \Pr(S_2\|B)\times U(B\|S_2) + \ldots + \Pr(S_n\|B)\times U(B\|S_n)$

  - Here, the probability of each state is calculated **conditional on a certain choice**. In short, we are treating our actions as evidence for the state most likely to obtain.
- Newcomb's problem specifies nothing about the relevant probability of either state and it is actually quite difficult to say anything (in isolation) about these probabilities. However, if the probability of each state is calculated conditional on certain choices, the following is certainly true.
  1. The probability of there being nothing in box 1 conditional on you taking the 2-box option is very high.
  2. The probability of there being £1,000,000 in box 1 conditional on you taking the 2-box option is very low.

3. The probability of there being nothing in box 1 conditional on you taking the 1-box option is very low.
4. The probability of there being £1,000,000 in box 1 conditional on you taking the 1-box option is very high.

- Let's assume, for simplicity, that "high probability" means 1 and "low probability" means 0. If so, calculating the expected utility of the two choices (using Nozick's method) yields the following result.
  - A = Nothing in box 1.
  - B = 1,000,000 in box 1.
  - C_1 = Take 1-box option.
  - C_2 = Take 2-box option.

1-Box Option

$$\begin{aligned} Exp(U(C_1)) &= \Pr(B|C_1)\times U(C_1|B) + \Pr(A|C_1)\times U(C_1|A) \\ &= (1 \times £1{,}000{,}000) + (0 \times £0) \\ &= £1{,}000{,}000 \end{aligned}$$

2-Box Option

$$\begin{aligned} Exp(U(C_2)) &= \Pr(B|C_2)\times U(C_2|B) + \Pr(A|C_2)\times U(C_2|A) \\ &= (0 \times £1{,}001{,}000) + (1 \times £1{,}000) \\ &= £1{,}000 \end{aligned}$$

- In other words, when expected utility calculations rely on probabilities of states that are conditional on choices, the 1-box options comes out as clearly superior.
- However, remember, if we calculate expected utility using unconditional probabilities, the expected utility of 2-boxing is higher than the expected utility of 1-boxing.
- Hence, as long as we don't conditionalize on the actions performed, the dominance principle is perfectly compatible with the MAXIMISE EXPECTED UTILITY rule.
- But, this of course raises the following question:
  - An act's influence on the probability of the state that we are in seems obviously relevant, so why **shouldn't** we conditionalize on actions performed?

Remember, it was proved earlier that dominating options always have the highest expected utility when this is calculated using unconditional probabilities.

10.2.1 Arguments for 2-Boxing

- Here are a two arguments for taking both boxes (2-boxing).
- Both arguments are essentially just reasons to think that the dominance rule must hold, and hence derivatively an argument against conditionalizing on actions performed when calculating expected utility.

Ask a Friend

- Suppose you're faced with Newcomb's problem but with the following twist.
  - You have a friend who can see into box 1. Now consider the following question: What advice would your friend give you?
  - Well, it seems that regardless of the state you are actually in, your friend would recommend that you take both boxes — for the following reasons.
    1. Either you are in state A, i.e. there is nothing in box 1, in which case your friend will see this and as a result recommend taking both boxes.
    2. Or you are in state B, i.e. there is a £1,000,000 in box 1, in which case your friend will see this and as a result recommend taking both boxes.
  - Plausibly when you know what advice your friend would give you regardless of the state you are in (and your friend is in a better epistemic position than you), you should take your friend's advice. Hence, you should take both boxes.
- Here is another way of making roughly the same argument.
  - Suppose you were offered a 'switching fee option', namely the possibility of paying £500 for the option of switching after your initial choice (and after looking what is in box 1).
  - This way, if you were to take just box 1 and realize that it is empty, you would be in a position to switch your choice to both boxes giving you a net profit of £500 (£1,000 – £500 fee).
  - However, clearly, there is no reason for you to do this. Since there is either £1,000,000 in box 1 or there isn't, you could just take both boxes to begin with. If you take both boxes, you will get the £1,000,000 if it is there and you won't if it is not. Switching will not make a difference.
  - Consequently, regardless of whether you pay the switching fee or not, taking 2 boxes is what you should do. This seems to suggest that 2-boxing is just the right option simpliciter.

Real Life Newcomb Problems

- Since Newcomb's problem is rather fantastic, it is worth considering cases that are more realistic to see if the same kind of asymmetry between dominance and expected utility (with probabilities conditional on actions) can arise in such cases.
  - Imagine we are in the following version of the prisoner's dilemma case. Two players, $\mathbf{P_1}$ and $\mathbf{P_2}$, can either cooperate or defect.
    1. If both players cooperate, each get 3 utils.
    2. If both players defect, each get 1 util.
    3. If $\mathbf{P}_i$ cooperates and $\mathbf{P}_j$ defects, $\mathbf{P}_i$ gets 0 utils and $\mathbf{P}_j$ gets 5 utils.

		$\mathbf{P_1}$	
		COOPERATE	DEFECT
$\mathbf{P_2}$	COOPERATE	3,3	0,5
	DEFECT	5,0	1,1

- This is effectively a problem in game theory, but if we assume that each player reasons in similar ways, this bears a close resemblance to Newcomb's problem.
- Notice that for both players there is a dominating option, namely to defect.
- Now suppose that conditional on $\mathbf{P_1}$ performing action A, there is 0.9 probability that $\mathbf{P_2}$ performs the same action.
- So, if we suppose that probabilities are conditional on choices, this yields the following expected utility for each player.

Cooperate (C)

$$\begin{aligned} Exp(U(C)) &= (0.9 \times 3) + (0.1 \times 0) \\ &= 2.7 \end{aligned}$$

Defect (D)

$$\begin{aligned} Exp(U(D)) &= (0.1 \times 5) + (0.9 \times 1) \\ &= 1.4 \end{aligned}$$

- In other words, calculating expected utility with probabilities conditional on actions yields a result which is inconsistent with the dominance rule — as in the case of Newcomb's problem.
- Sadly, this is not particularly helpful, because it is not totally clear what the right decision is in a prisoner's dilemma case.

Keep Smoking?

- Let's consider a different, potentially more helpful case.
    - Suppose, for the sake of argument, that smoking does not cause cancer, but rather that some people are genetically predisposed to getting cancer.
    - However, suppose further that a genetic predisposition to cancer also causes a disposition to smoke.
    - Hence, cancer and smoking are merely correlated because they have a common cause, namely a genetic predisposition — but they are probabilistically dependent, i.e. $\Pr(\textit{Get-Cancer}|\textit{Smoking}) > \Pr(\textit{Get-Cancer})$.
    - If these are the facts (and this is the assumption), then if you have a desire to smoke, you should just keep smoking, even if you also have a desire to avoid cancer.

- However, we will only predict this if we assume that the probabilities of each state obtaining are **not** conditional on the actions performed!
- Consider the following decision table.

		A	B
		Cancer	*No Cancer*
S	**Smoke**	1	6
$\neg S$	**Don't Smoke**	0	5

- The assumption here is that not getting cancer while smoking is most desirable to you.
- Since smoking is evidence of a genetic disposition to cancer, this raises the overall chance of getting cancer. Hence, we'll assume the following.
  - $\Pr(A|S) = 0.8$
  - $\Pr(A|\neg S) = 0.2$
- This yields the following expected utilities.

Smoke

$$\begin{aligned} Exp(U(S)) &= (0.8 \times 1) + (0.2 \times 6) \\ &= 2 \end{aligned}$$

Don't Smoke

$$\begin{aligned} Exp(U(\neg S)) &= (0.2 \times 0) + (0.8 \times 5) \\ &= 4 \end{aligned}$$

- Hence, the recommendation here is to not smoke — again violating dominance.
- This looks like the wrong prediction. After all, while smoking is **evidence** that you might have a genetic disposition, simply refraining from smoking will not make it any less likely — either you have the genetic disposition or you don't.
- In a sense, deciding not to smoke seems like denying yourself a good outcome only to avoid getting (the inevitable) bad news.
- This seems like a compelling argument for not calculating expected utilities on the basis of probabilities conditional on actions performed.

A Response

- One response to this problem is to argue that in the above case the genetic disposition must cause a desire to smoke.
- So, if the agent has a desire to smoke (and is aware of this), then the fact that he actually smokes does not provide any further evidence that he has cancer (in addition to the evidence already coming from having the desire).

- With this knowledge in hand, the probability of each state (getting cancer vs. not getting cancer) is the same conditional on smoking. If the agent has a desire to smoke, then there is (let's suppose) 0.8 probability of the agent getting cancer regardless of the smoking — and as a result the calculation of expected utility should proceed as follows.

Smoke

$$\begin{aligned} Exp(U(S)) &= (0.8 \times 1) + (0.2 \times 6) \\ &= 2 \end{aligned}$$

Don't Smoke

$$\begin{aligned} Exp(U(\neg S)) &= (0.8 \times 0) + (0.2 \times 5) \\ &= 1 \end{aligned}$$

- Hence, the advocate of conditionalizing on actions performed can make the correct prediction in the smoking case too.
- Of course, this prediction does rely on two somewhat contentious assumptions.
  a. In between the actual smoking and the genetic disposition, there has to be a **desire** to smoke — but the case could possibly be reconstructed to avoid this.
  b. Agents must always know what they desire.

10.3 Causal vs. Evidential Decision Theory

- The two approaches to calculating expected utilities corresponds to two different strands of decision theory, namely causal decision theory and evidential decision theory.
  - **Causal Decision Theory**: Rational agents try to maximise **causal** expected utility.
- In causal decision theory, the expected utility of an action A is calculated on the basis of unconditional probabilities, cf. below.

$$Exp(U(A)) = \Pr(S_1)\times U(A|S_1) + \Pr(S_2)\times U(A|S_2) + \ldots + \Pr(S_n)\times U(A|S_n)$$

- Here, expected utility is a measure over the outcomes that you can expect to bring about by performing some action.
- In contrast:
  - **Evidential Decision Theory**: Rational agents try to maximise evidential expected utility.
- So, in evidential decision theory, the expected utility of an action A is calculated on the basis of conditional probabilities

$$Exp(U(A)) = \Pr(S_1|A)\times U(A|S_1) + \Pr(S_2|A)\times U(A|S_2) + \ldots + \Pr(S_n|A)\times U(A|S_n)$$

- Here, expected utility is measure over the kind of results your action would be evidence for.

10.3.1 Arguments for Evidential Decision Theory

Broken Windshields

- Remember the case from Joyce that we considered earlier:

> Suppose you have just parked in a seedy neighborhood when a man approaches and offers to "protect" your car from harm for $10. You recognize this as extortion and have heard that people who refuse "protection" invariably return to find their windshields smashed. Those who pay find their cars intact. You cannot park anywhere else because you are late for an important meeting. It costs $400 to replace a windshield. Should you buy "protection"? **Dominance** says that you should not. Since you would rather have the extra $10 both in the even that your windshield is smashed and in the event that it is not, **Dominance** tells you not to pay.
>
> Joyce (1999, 115-116)

- This case had the following decision table.

		B *Windshield Broken*	**¬*B*** *Windshield Intact*
P	**Pay Insurance**	–$410	–$10
¬*P*	**Don't Pay Insurance**	–$400	$0

- As Joyce points out, dominance instructs you not to pay insurance, and since dominance is never violated in causal decision theory, it follows that causal decision theory predicts that a rational agent should not pay for insurance — which intuitively is the wrong result.
- Here, evidential decision theory has the advantage, because if expected values are calculated using conditional probabilities, it is easy to predict that you should pay for insurance.
- For example, let's suppose that the probability of a broken windshield **conditional on paying** is 0.05 and that the possibility of an broken windshield **conditional on not paying** is 0.95.
- This yields the following expected values.

Pay Insurance

$$\begin{aligned} Exp(U(P)) &= (0.05 \times -\$410) + (0.95 \times -\$10) \\ &= -\$30 \end{aligned}$$

Don't Pay Insurance ($\neg P$)

$$\begin{aligned} Exp(U(\neg P)) &= (0.95 \times -\$400) + (0.05 \times 0) \\ &= -\$380 \end{aligned}$$

- So, evidential decision theory seems to make the correct prediction here.

Newcomb's Problem

- While it is controversial what the correct decision is in Newcomb's problem, it seems that people who take one box (i.e. follows the recommendation of evidential decision theory) tend to end up with a better outcome (i.e. walk away with more money).

Chapter 11

Framing Effects

11.1 Risk Aversion

- It is well known that people are generally adverse to risk. This becomes particularly clear when studying preferences with regards to certain gambles. Here is an example.
  - When given a choice between *A* and *B* below, most people choose *B*.

    A : 85% chance of winning $1,000 (15% chance to win nothing).

    B : 100% chance of winning $800.
  - Notice that the expected value of each choice is higher in *A* than in *B*. The expected value of *B* is $800, but the expected value of *A* is (0.85×$1,000) + (0.15×0) = $850.
  - This is often characterized as a instance of risk aversion, namely a preference for a sure gain over a gamble with higher expected value.
- One famous proposal to explain risk aversion was put forward by Daniel Bernoulli. According to Bernoulli, people do not evaluate prospects in terms of monetary expectation, but rather in terms of **subjective value** (or **utility**).
- Bernoulli suggested that subjective value is a concave function of money. So, for example, given such a function, the difference in subjective value (the utility) between $100 and $200 is greater than the difference in subjective value between $1,100 and $1,200.
- Thus, the subjective value of a sure gain of $800 is higher than the subjective value of a gamble with 80% chance to win $1,000 even though these have the same expected monetary value.

11.1.1 Gains vs. Losses

- Outcomes of decisions are often described in terms of total wealth. For example, a $20 bet on the toss of a fair coin is often described in the following way:

Wealth+$20   or   *Wealth*–$20

- However, from a psychological perspective, this is somewhat unrealistic. People do not generally think of relatively small outcomes in proportion to their total wealth.
- According to Kahneman and Tversky (K&T), agents generally think of gambles in terms of **gains** and **losses**. Moreover, K&T have conducted multiple studies which show that whether a gamble is conceived of as a potential gain or a potential loss makes a significant difference to the agent's decision.
    - **The Subjective Value of Gains**
      The subjective value of a (monetary) **gain** appears to be a concave function, i.e. as mentioned above, the difference in subjective value (the utility) between $100 and $200 is greater than the difference in subjective value between $1,100 and $1,200.
    - **The Subjective Value of Losses**
      However, the difference in subjective value between a **loss** of $200 as opposed to a loss $100 seems greater than the difference in subjective value between a loss of $1,200 and a loss $1,100. Hence, the subjective value of a monetary **loss** appears to be a convex function.
- This hypothetical value function thus has the S-shape illustrated below:

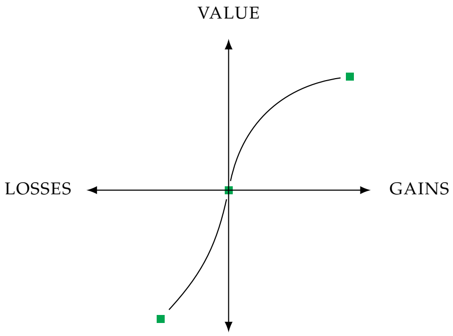


FIGURE I: HYPOTHETICAL VALUE FUNCTION

- The function is considerably steeper in the domain of losses. This is meant to capture that people are generally more **loss averse** than they are **gain attracted**. In other words, a loss of $*x* is generally more aversive than a gain of $*x* is attractive.
- Among other things, this hypothesis can provide an explanation for a study conducted by K&T (1984) where most respondents in a sample of undergraduates refused to bet $10 on the toss of a fair coin unless they stood to win at least $30.

Risk Seeking

- Notice that because the value of losses is a convex function, this means that people generally become more risk seeking the higher the already guaranteed loss is.

- This is evidenced by a study where of the two options *A* and *B* below, the majority of the respondents chose *A*.

  A : 85% chance to lose $1,000 (15% chance to lose nothing).

  B : 100% chance to lose $800.

- The expected monetary value of these gambles is higher in *B* where the expected monetary value is –$800 as opposed to *A* where the expected monetary value is –$850.

Subjective Values and Rationality

- A question that might be raised now is whether it is wrong to be risk averse in the domain of gains and risk seeking in the domain of losses? I.e. is it irrational to have these preferences?
- According to expected utility theory, the answer seems to be affirmative.
- The axioms of expected utility theory are supposed to describe a range of necessary conditions for rationality. This includes e.g. the axioms below.

  - **Dominance**
    When an option *A* dominates an option *B*, the expected utility *A* is always higher than the expected utility of *B*. Hence, one should never choose dominated options.
  - **Transitivity**
    If the expected utility of *A* is higher than the expected utility of *B*, and the expected utility of *B* is higher than the expected utility of *C*, then the expected utility of *A* is higher than the expected utility of *C*. Hence, in such cases, an agent should always choose *A* over *C*.
  - **Invariance** (Completeness)
    An agent always (i) prefers *A* over *B*, (ii) prefers *B* over *A*, or (iii) is indifferent about *A* and *B*. This means that if two versions of one single decision problem can be shown to be equivalent, the decision problem should elicit the same preferences independently of the version that is being considered.

- The problem is that the behavior documented in studies by K&T leads to violations of these axioms. To appreciate why, let's consider **invariance**.

11.2 Outcome Framing

- The following two problems are from a study by K&T (1984) on decisions under risk. The %-numbers indicate the proportion of people who chose that option.

PROBLEM I

Imagine that the US is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows:

- If Program A is adopted, 200 people will be saved (72%).
- If Program B is adopted, there is one third probability that 600 people will be saved and a two-thirds probability that no people will be saved (28%).

Which of the two programs would you favor?

- Notice that the potential outcomes are framed here in terms of lives saved. Since "lives saved" is naturally taken as a gain, then the majority of the respondents (72%) chose the **risk averse** option.
- Now consider the following contrast.

PROBLEM II

Imagine that the US is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows:

- If Program C is adopted, 400 people will die (22%).
- If Program D is adopted, there is a one-third probability that nobody will die and a two-thirds probability that 600 people will die. (78%).

Which of the two programs would you favor?

- Here the potential outcomes are framed in terms of lives lost, i.e. a potential loss. As a result, the majority of the respondents (72%) had a risk seeking preference in this case.
- The point here is that with respect to the objective outcome, PROBLEM I and PROBLEM II are equivalent.
- Even more interesting, when the respondents were made aware of their conflicting answers, they often still retained their judgments after rereading the cases.
- This result (and other similar results) demonstrates a somewhat serious problem, namely that what appears to people to be the intuitively correct decision depends essentially on the way in which the problem is **framed**. And this then raises a fundamental question about the terms with which various problems should be described.

- This result also suggests that if one wants to test the robustness of an agent's preferences, one should try to frame the decision problem in more than one way and check for consistency.

11.2.1 The Psychophysics of Chances

- On the standard picture of expected utility, the expected utility of an action A is calculated as follows:

$$Exp(U(A)) = \Pr(S_1)\times U(A|S_1) + \ldots + \Pr(S_i)\times U(A|S_i)$$

- Simplifying a bit, suppose you are given a lottery ticket for a lottery that pays \$300 to the winner — the value of the ticket should then simply vary as a function of the probability of the ticket winning, i.e.
  - If the ticket has 0.5 chance of winning, the ticket is worth \$150.
  - If the ticket has a 0.01 chance of winning, the ticket is worth \$3.
- More generally,
  - If the $\Pr(win) = x$, the ticket is worth $\$300\times x$.

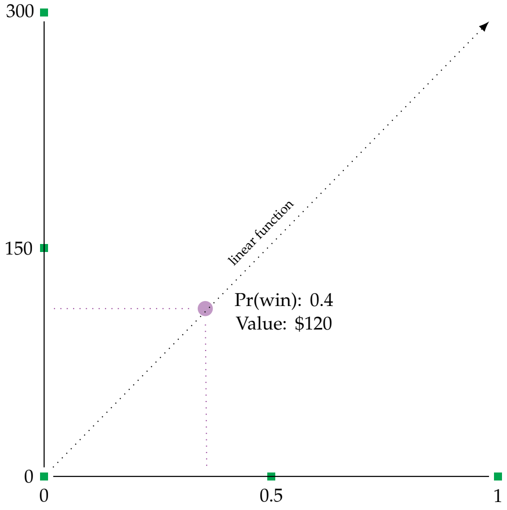


FIGURE II: LINEAR FUNCTION

- However, studies show that the perceived value of the function is not linear — for example an increase from 0% to 5% chance of winning generally seems more significant than an increase from 35% to 40%.
- The general pattern is as follows:

- Small increases in the probability of winning by e.g. 5% generally increases the value of the prospect less than 5%.
- However, small increases in the probability of winning at the endpoints generally increases the value of the prospects by more than 5%.

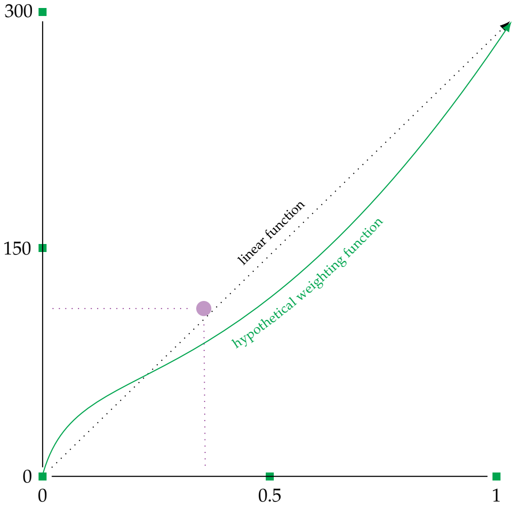


FIGURE III: HYPOTHETICAL WEIGHTING FUNCTION

- This general tendency to underweight moderate to high probabilities relative to sure things contributes to risk aversion in gains since it reduces the attractiveness of gambles.
- Correspondingly underweighting moderate to high probabilities relative to sure things contributes to risk seeking in losses.
- This non-linearity of decision weights now leads to violations of invariance.

PROBLEM III

Consider the following two-stage game. In the first stage, there is a 75% chance to end the game without winning anything and a 25% chance to move into the second stage. If you reach the second stage you have a choice between:

- **A** : A sure win of $30 (74%).
- **B** : 80% chance to win $45 (26%).

Your choice must be made before the game starts, i.e. before the outcome of the first stage is known. Please indicate what option you prefer.

- Now consider the problem below.

> PROBLEM IV
> Which of the following options do you prefer:
>
> - **C** : 25% chance to win $30 (42%).
> - **D** : 20% chance to win $45 (58%).

  - In PROBLEM I, there is a 0.25 chance of moving into the second stage of the game, and hence there is a 0.25 chance of winning $30 dollars. So, **A** and **C** are equivalent.
  - In PROBLEM II, there is a 0.25×0.8 = 0.2 chance of winning $45, so **B** and **D** are equivalent.

- Since the two problems are equivalent, an agent violates **invariance** if her answer for either of the two problems differ.
- According to K&T, this result can be explained in terms of two factors:
  - The framing of the decision problem.
  - The non-linearity of decision weights.
- For example, because of the **framing** of PROBLEM III, people simply ignore the first phase and then focus their attention on what happens in the second.
- Focusing only on the second part of the game, the agent is facing a sure gain of $30 or an 80% chance of winning $45. Since sure things are overweighted in comparison with moderate or high probabilities, the sure gain is considered more attractive.

11.2.2 Normative vs. Descriptive Projects

- So, what should we make of K&T's observations? Is this a problem for expected utility theory? And if so, in which sense is it a problem?
- There are (at least) two ways to construe the general aim of expected utility theory, namely as:
  - A normative account (of rationality).
  - A descriptive account (of rationality).
- If one construes the overall theory as a normative account, i.e. an account of what people **should** do when faced with various decision problems in order to count as rational, it is not obvious that these empirical findings are problematic. Or at the very least, it is debatable. Whether people are in fact being irrational when they violate the principles of expected utility theory is a substantial philosophical question.

- However, if one construes the overall theory as a descriptive account, i.e. an account of what people **actually** do when faced with decision problems, K&T's observations seem to pose a serious problem.
- The empirical findings clearly show that people do not always respect the axioms of expected utility theory (and hence the axioms of probability theory). So, if expected utility theory is supposed to be used as a model for predicting people's choices in various situations, it is going to make many incorrect predictions.

Challenges for Descriptive Analyses: Implicit Bias

- People make reflective and balanced decisions all the time, but it is well known that people are not generally particularly proficient at mathematics, logic or probability theory.
- However, studies show that even people who are trained in e.g. mathematics, statistics, and probability theory are prone to making fairly basic mistakes when calculating probabilities. One explanation for the occurrences of these basic mistakes is so-called implicit biases.
- While we do not have time to cover this topic in any detail, we will end by doing a few exercises and discussing the results.

References

Allais, Maurice 1953. 'Le comportement de l'homme rationnel devant le risque: critique des postulats et axioms de lécole Américaine'. *Econometrica*, 31, 4: 503–546.

Barwise, Jon and Etchemendy, John 1999. *Language, Proof and Logic*. CSLI Publications.

Joyce, James M. 1999. *Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press.

Kahneman, Daniel and Tversky, Amos 1984. 'Choices, Values, and Frames'. *American Psychologist*, 39, 4: 341–50. Reprinted in Kahneman and Tversky (2000).

Kahneman, Daniel and Tversky, Amos (eds.) 2000. *Choices, Values, and Frames*. Cambridge: Cambridge University Press.

Sider, Theodore 2010. *Logic for Philosophy*. Oxford: Oxford University Press.

Weatherson, Brian 2011. 'Logic of Decision'. Lecture Notes (unpublished ms.).