CY0002 Ethics Notes

Hankertrix

April 29, 2025

Contents

1 Definitions

1.1 Reasoning

Reasoning is the process by which one draws **conclusions** from a set of premises. Reasoning or inference may be represented as follows:

$$\phi \vdash_X^M \psi$$

Where:

- ϕ is the set of premises
- $\vdash$ represents that ψ is provable from ϕ
- M refers to the mode of inference, which can be deductive, inductive, abductive, analogical, etc.
- X is the inferential mechanism.
- ψ is the conclusion set.

1.2 Inductive reasoning

Inductive reasoning refers to making observations to find patterns and using the patterns to reason about things. Inductive reasoning usually works better with a large sample size.

1.3 Deductive reasoning

Deductive reasoning refers to drawing conclusions using a formal logic system, like mathematics, for example.

1.4 Abductive reasoning

Abductive reasoning refers to seeking the simplest and most likely conclusion from a set of observations. Abductive reasoning is usually used when there is a small sample size.

1.5 Analogical reasoning

Analogical reasoning is a special type of inductive reasoning where perceived similarities are used as a basis to infer some further similarity that has not been observed yet.

1.6 Modus ponens

- Modus ponens, also known as modus ponendo ponens, which is Latin for "mode that by affirming affirms".
- It can be summarised as "P implies Q. P is true. Therefore, Q must also be true."

1.7 Natural deduction (Gentzen-style logic system)

Natural deduction is a kind of proof calculus in which logical reasoning is expressed by rules that are closely related to the "natural" way of reasoning.

1.8 Hilbert-style logic system

- A Hilbert-style proof system is a type of formal proof system.
- It is defined as a deductive system that generates theorems from axioms and inference rules, especially if the only inference rule is modus ponens.
- Every Hilbert system is an axiomatic system.

1.9 Propositional logic (PL)

$$\phi \vdash_{PL}^{ND} \psi$$

Where:

- ϕ is the set of premises
- $\vdash$ represents that ψ is provable from ϕ
- ND refers to the deductive mode of natural deduction
- PL refers to the inferential mechanism of propositional logic
- ψ is the set of conclusions derived from the set of premises

1.10 Proposition

A proposition is a statement or an assertion that can be either true or false.

1.11 Propositional variable

A propositional variable is a variable that is used to capture the content of a proposition, which is a statement that can be either true or false.
Usually they are p and q.

1.12 Antecedent

The antecedent is the statement in which a statement is inferred. It is p in the example below:

$$p \rightarrow q$$

1.13 Consequent

The consequent is the statement that is inferred from another statement. It is q in the example below:

$$p \rightarrow q$$

1.14 Syllogism

A syllogism is a kind of logical argument that applies deductive reasoning to arrive at a conclusion based on two propositions that are asserted or assumed to be true.

1.15 Quantificational or first-order predicate logic (QL)

$$\phi \vdash_{QL}^{ND} \psi$$

Where:

- ϕ denotes any well-formed formula (wff) in quantificational logic
- $\vdash$ represents that ψ is provable from ϕ
- ND refers to the deductive mode of natural deduction
- QL refers to the inferential mechanism of quantificational or first-order predicate logic
- ψ is the set of conclusions derived from the set of premises

1.16 Deductive argument assumption

The deductive argument assumption assumes that the conclusion of an argument cannot contain more information than is held in its premises.

1.17 Descriptive proposition ("is")

A descriptive proposition is a statement of fact.

1.18 Normative proposition ("ought")

A normative proposition is a proposition that contains a value judgment, like a moral judgment or ethical judgment.

1.19 Hume's Law (Autonomy of Ethics / NOFI principle)

Hume's First Law states that we cannot deduce how things ought to be or what ought to be done, which is a moral judgment from how things are, which is a statement of fact. This is also known as the view that ethics is autonomous. This is also known as the no ought from is principle, or NOFI.

1.20 Russell's Law

You can never arrive at a general proposition by inference from particular propositions alone. You will always have to have at least one general proposition in your premise.

$$\phi \nvdash \psi$$

Where:

- ϕ is the particular proposition
- $\nvdash$ means "does not entail that"
- ψ is the universal or general proposition

1.21 Hume's Second Law (The Problem of Induction)

Hume's second law states that you cannot derive propositions about the future from propositions about the past or present.

$$\phi \nvdash \psi$$

Where:

- ϕ are the propositions about the past or present
- $\nvdash$ means "does not entail that"
- ψ are the propositions about the future

1.22 Kant's Law

Kant's Law states that you cannot derive necessary propositions from propositions about the actual world.

$$\phi \nvdash \psi$$

Where:

- ϕ are the propositions about the actual world
- $\nvdash$ means "does not entail that"
- ψ is the necessary propositions

1.23 Barrier construction theorem

Implication barrier	Topic	Description
Hume's (1739/40) Law	Normativity	You cannot derive **normative propositions** (Op) from **descriptive propositions** (p)
Russell's (1918) Law	Generality	You cannot derive **general propositions** ($\forall xFx$) from **particular propositions** (Fa)
Hume's (1748) Second Law	Time	You cannot derive **propositions about the future** (Fp) from **propositions about the past or present** (Pp)
Kant's (1787) Law	Necessity	You cannot derive **necessary propositions** ($\Box p$) from **propositions about the actual world** (p)

1.24 Geach-style conditionalisation

Geach-style conditionalisation refers to embedding "ought" propositions in conditionals, which appear to allow us to derive valid is-ought inferences.

1.25 A priori

A priori is a Latin phrase meaning "from the earlier".

1.26 A posteriori

A posteriori is a Latin phrase meaning "from the later".

1.27 Denotatum (plural: denotata)

Denotatum means a denotation of a word or an expression. The denotation of a word or expression is its strictly literal meaning, so the English word "warm" would denote a high temperature.

1.28 Argumentum a fortiori (a fortiori)

Argumentum a fortiori is a Latin phrase meaning "argument from the strong reason".

1.29 Shew / Shewn

Shew is just an archaic alternative form of show.

1.30 Counterfactual conditionals

Counterfactual conditionals are conditional sentences which discuss what would have been true under different circumstances, e.g., "If Peter believed in ghosts, he would be afraid to be here."

1.31 Reductio ad absurdum

Reductio ad absurdum, Latin for "reduction to the absurdity", disproves a proposition by showing that it leads to absurd or untenable conclusions.

1.32 Contradiction ($\perp$)

A contradiction is a statement that is always false.

1.32.1 Principle of explosion

Anything follows from a contradiction (including a **normative proposition** of the form Op).

1.33 Tautology ($\top$)

- A tautology is a statement that is always true.
- A tautology or logical **truth** follows from anything (including a **descriptive proposition** of the form p).

1.34 Enthymemes

Enthymemes are arguments with hidden premises.

1.35 Salva Veritate (Salva Validitate)

Salva Veritate is a Latin phrase for "with unharmed truth". It means that something can be done without changing the validity of the argument.

1.36 Contingent truth

A contingent truth is true as it happens, or as things are, but that did not have to be true.

1.37 Ampliative

Ampliative means "extending" or "adding to that which is already known".

1.38 Nash equilibrium

Nash equilibrium refers to a play in which each strategy is the **best response** to the strategy played by the other person.

1.38.1 Example

		P2	
		STRATEGY 3	STRATEGY 4
P1	STRATEGY 1	(5, 5)	(2, 3)
	STRATEGY 2	(0, 1)	(4, 2)

- P1's strategy 2 is the **best response** to P2's strategy 4 (and vice versa).
- P1's strategy 1 is the **best response** to P2's strategy 3 (and vice versa).

The cells coloured in **yellow** denote the **Nash equilibria**.

1.39 Pareto optimality (Pareto efficiency)

A state of affairs such that there is **no alternative state of affairs** that would **make some people better off without making at least one person worse off**.

1.39.1 Example

		P2	
		STRATEGY 3	STRATEGY 4
P1	STRATEGY 1	(5, 5)	(2, 3)
	STRATEGY 2	(0, 1)	(4, 2)

The cell coloured in **green** denotes a **Pareto-optimal** state of affairs.

1.40 Validity of an argument

An argument is **valid** if, **assuming the truth of all its premises**, its conclusion must, by **logical necessity, be true too**.

1.41 Soundness of an argument

An argument is sound if **all of its premises are in fact true**, or it does not contain any **false premises**, and it is a **valid argument**.

1.42 Classificatory moral commitments

Classificatory moral commitments are defined as the commitments that result from delineating the **scope of the moral domain**.

1.43 Substantive moral commitments

Substantive moral commitments are defined as a **normative bias**.

1.44 Normative neutrality

- Normative neutrality between **competing moral standards** and **rules of conduct** is where the cut between **metaethics (2^{nd}-order theory)** and **normative theory (1^{st}-order theory)** is made.
- Essentially, normative neutrality is what separates **metaethics and normative theory**.

1.45 Axiology

Axiology just means value theory.

1.46 Reflective equilibrium

Reflective equilibrium is a method of balancing moral principles and judgments to arrive at the content of justice.

1.47 Eudaimonia

- Eudaimonia is a Greek word that means the state or condition of good spirit, and is often translated as happiness or welfare. In Aristotle's works, it means the highest human good.
- It is a certain **flourishing** or the **sort of happiness worth seeking or having**.

1.48 Hedonism

1.48.1 Psychological hedonism

Only **pleasure (happiness)** or **pain (unhappiness)** motivates us.

1.48.2 Ethical hedonism

Only **pleasure (happiness)** has **value** and only **pain (unhappiness)** has **disvalue**.

1.49 Avant la lettre

Avant la lettre means that a concept exists even before a term is coined for it.

1.50 Elenchus elenctic (The Socratic method)

The Elenchus elenctic is a form of argumentative dialogue between individuals based on asking and answering questions.

1.51 Agent

The agent is the person who is performing an action.

1.52 Patient

The patient is the person on whom the action is performed.

1.53 Prima facie

Prima facie is a Latin phrase meaning "at first sight", or "based on first impression". It is used in philosophy to indicate that something is sufficient or plausible unless rebutted.

1.54 Virtue ethics

Virtue ethics is concerned with providing an account of the **moral virtues**.

1.55 Virtual epistemology

Virtue epistemology is concerned with providing an account of the **intellectual virtues**.

2 Prior's paradox

2.1 Assumptions

Assumption	Description
Dichotomy assumption (A1)	All propositions may be categorised as either **ethical or non-ethical**
Deductive argument assumption (A2)	The **conclusion** of an argument cannot contain more information than its **premises**.

2.2 Proposition

Either tea drinking is common in England, or it ought to be the case that all New Zealanders are shot, formalised as $p \vee Oq$. According to the dichotomy assumption, the proposition is either **ethical** or **non-ethical**.

Horn 1	Horn 2
$p \vee Oq$ is ethical.	$p \vee Oq$ is non-ethical.
If $p \vee Oq$ is **ethical**, then: P1 (non-ethical): Tea drinking is common in England. C (ethical): Therefore, either tea drinking is common in England, or it ought to be the case that all New Zealanders are shot.	If $p \vee Oq$ is **non-ethical**, then: P1 (non-ethical): Either tea drinking is common in England, or it ought to be the case that all New Zealanders are shot. P2 (non-ethical): Tea drinking is not common in England. C (ethical): Hence, it ought to be the case that all New Zealanders are shot.

2.2.1 Dilemma

Whether we accept horn 1 or horn 2, we make **is-ought inferences** that are perfectly **valid**.

Horn	Classification of $p \vee Oq$	Premise set	Conclusion set
Horn 1: $p \vdash p \vee Oq$	$p \vee Oq$ is **ethical**	**Non-ethical**	**Ethical**
Horn 2: $p \vee Oq, \neg p \vdash Oq$	$p \vee Oq$ is **non-ethical**	**Non-ethical**	**Ethical**

Prior's paradox is a **dilemma without escape**. Since **Hume's Law**, even as a **one-way implication barrier**, is violated in every possible instance, it must be **false**. Hence, ethics is not **logically autonomous**.

3 Defending Hume's Law

1. Admit the **converse of the is-ought thesis**.

2. Exclude **contradictions** from the premise set ϕ. ϕ should be defined as a **consistent or contradiction-free** set of **descriptive propositions**.

3. Exclude **tautologies** from the conclusion set ψ. ψ should be defined as a **normative proposition** that is not already logically true.

4. Rule out **enthymematic arguments**. When the hidden premises of **enthymemes** are restored, the **premise set** ϕ will have at least one **normative proposition**. Hence, these arguments will no longer be obvious counterexamples to Hume's Law.

5. Concede the **contraposition with "ought" implies the "can"** case. We should concede that **"cannot" implies "not obligatory"** yields a **special case** in which $\phi \vdash \psi$.

6. Rule out **Geach-style conditionals** as non-ethical propositions. It gives rise to embedded "ought" propositions of the form $Op \rightarrow Oq$.

   With $Op \rightarrow Oq$, "ought" statements Op, Oq are being embedded into more complex logical structures, but there is no commitment to the truth or falsity of either Op or Oq.

7. Rule out **mixed propositions** from the premise set ϕ and the conclusion set ψ. We can replace the **dichotomy** assumption with the **trichotomy** assumption, where all propositions may be categorised as either **ethical, non-ethical, or mixed**.

3.1 Issues with move 7

Mixed propositions have an indispensable role in ethical reasoning and argumentation. Purely normative propositions are rarely encountered in the real world, outside the philosopher's laboratory.

Examples include:

Proposition in natural language	Formal representation
If you refrain from helping the old lady across the road, then you ought to be blamed.	$\neg p \rightarrow Oq$
Either you help the old lady across the road, or you ought to be blamed for not doing so.	$p \vee Oq$
It is necessarily the case that if p, then it is obligatory that q.	$\Box(p \rightarrow Oq)$
It is necessarily the case that for all x, then Fx, then it is obligatory that Gx.	$\Box\forall x(Fx \rightarrow Gx)$

3.2 Gerhard Schurz substitution

- If a **mixed conclusion** ϕ is derivable from a **purely non-ethical premise set** ϕ, then ψ is completely **O-irrelevant**.
- Apply the **O-restricted propositional substitution function** σ.
- Substitute r (any proposition whatsoever) for q on exactly those occurrences of q outside the scope of O, i.e.

  p (non-ethical) $\vdash p \vee Oq$, (mixed) $\xrightarrow{\text{Apply } \sigma}$ p (non-ethical) $\vdash p \vee Or$ (mixed)

- The **O-restricted substitution** (σ) can be made without compromising the validity of the argument; hence, the mixed conclusion $p \vee Oq$ is completely **O-irrelevant** relative to the premise set.
- If a **mixed premise set** ϕ is used to derive a **purely ethical conclusion** ψ, then ϕ is completely **is-irrelevant**.
- Apply the **is-restricted propositional substitution function** σ'.
- Substitute r (any proposition whatsoever) for p on exactly those occurrences of p outside the scope of O, i.e.

  $p \vee Oq, \neg p$ (mixed) $\vdash Oq$ (ethical) $\xrightarrow{\text{Apply } \sigma'}$ $r \vee Oq, \neg r$ (mixed) $\vdash Oq$ (ethical)

- The **is-restricted substitution** (σ') can be made without compromising the validity of the argument; hence, the **mixed premise set** $p \vdash Oq, \neg p$ is completely **is-irrelevant** relative to the conclusion.

3.3 Gibbard-Karmo-Singer semantics

Gibbard-Karmo-Singer semantics is just a way of determining whether a set of propositions will result in ethical conclusions or not. It works like this:

1. Consider the truth value of the propositions in a possible world, such as the actual world we live in.
2. Consider the truth value of the propositions and conclusions in an ethical standard.
3. Swapping the ethical standard for another ethical standard without changing the world.
4. If the truth value of the conclusions changes when you change the ethical standard, like the conclusions change from true to false, then the conclusions are ethical.
5. Otherwise, the conclusions are non-ethical, because the ethical standard being used is not relevant to the truth value of the conclusions.
6. If the conclusions are non-ethical, and you want to figure out which possible worlds the set of propositions will result in ethical conclusions, swap out the world for another one and repeat steps 2 to 5.

3.4 Shorter's position

- The **conclusion** of an argument may be of **some importance (ethically speaking)** in deriving certain moral duties only if it is arrived at in some other way than employing an **is-ought inference**.
- Hence, the **is-ought** inference is **not of importance (ethically speaking)**.
- We need to distinguish between the **seriousness of the conclusion arrived at (ethically speaking)** and the **seriousness of the is-ought inference** by which the conclusion is arrived at.

3.4.1 Tea drinking example

- P1 (non-ethical): Tea drinking is common in England.
- C (ethical): Therefore, either tea drinking is common in England or it ought to be the case that all New Zealanders are shot.

Step	Description
1	P1 is either true or false.
2	If P1 is true, then the **is-ought inference** lends support to C. However, C will be of no help or use to us in deriving certain moral duties. If P1 is false, then the **is-ought inference** lends no support to C. However, if P1 is false and C is true, then we can derive the duty to shoot all New Zealanders, and C may be of some importance (ethically speaking).
3	Therefore, whether P1 is true or false, the **is-ought inference** is **useless**. It either **renders C ethically useless insofar as it supports C (when P1 is true) or does not support C** (when P1 is false)

3.4.2 Undertaker example

- P1: Undertakers are church officers.
- C: Therefore, undertakers ought to do whatever all church officers ought to do.
- P2: All church officers ought to ϕ.

Step	Description
1	The **is-ought inference** either makes do without P2 or incorporates P2.
2	If the **is-ought inference** makes do without P2, then it lends support to C. However, C will be **useless** without P2 as an undertaker can only derive a **concrete moral duty** with both P1 and P2. If the **is-ought inference** incorporates P2, then it lends support to C' (All undertakers ought to ϕ rather than C (Undertakers ought to do whatever all church officers ought to do). Therefore, C will become **useless** with P2.

4 Modal concepts

Mode	Domain	Categories	Logical state of play in 1951
Mode 1 (Alethic)	Truth	Necessary, Possible, and Contingent	**Alethic modal logic** with the **modal operators** $\Box$ and $\Diamond$
Mode 2 (Epistemic)	Knowledge	Verified, Falsified, and Indeterminate	Minimal logical treatment
Mode 3 (Deontic)	Actions	Obligatory, Permissible, and Impermissible	Minimal logical treatment
Mode 4 (Existential)	Existence	Universal, Existential, and Empty	First-order predicate or quantificational logic (QL) with the quantifiers $\forall$ and $\exists$

5 Deontic logic

Deontic logic is a field of philosophical logic that is concerned with obligation, permission, and related concepts. The word "deontic" comes from the Greek word "deon" which means "that which is binding or proper".

5.1 Analogies between the deontic mode and the alethic mode

5.1.1 Analogy 1

There are 2 operators.

- Alethic mode

$$\Box p \stackrel{\text{def}}{=} \neg\Diamond\neg p$$
$$\Diamond \stackrel{\text{def}}{=} \neg\Diamond\neg p$$

  $\Box$ and $\Diamond$ are De Morgan duals.

- Deontic mode

$$Op \stackrel{\text{def}}{=} \neg P\neg p$$
$$Pp \stackrel{\text{def}}{=} \neg O\neg p$$

  O and P are De Morgan duals.

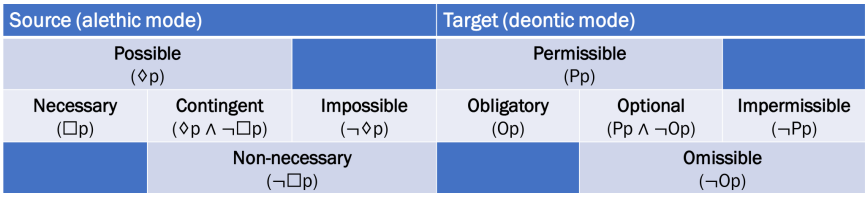
5.1.2 Analogy 2

There are **5 statuses** that can be defined in terms of **2 operators**.

Source (alethic mode)	Target (deontic mode)
It is **necessary** that p ($\Box p$)	It is obligatory that p (Op)
It is **possible** that p ($\Diamond p$)	It is **permissible** that p (Pp)
It is **impossible** that p ($\neg\Diamond p$)	It is **impermissible** that p($\neg Pp$)
It is **non-necessary** that p ($\neg\Box p$)	It is **omissible** that p($\neg Op$)
It is **contingent** that p ($\Diamond p \wedge \neg\Box p$)	It is **optional** that p($Pp \wedge \neg Op$)

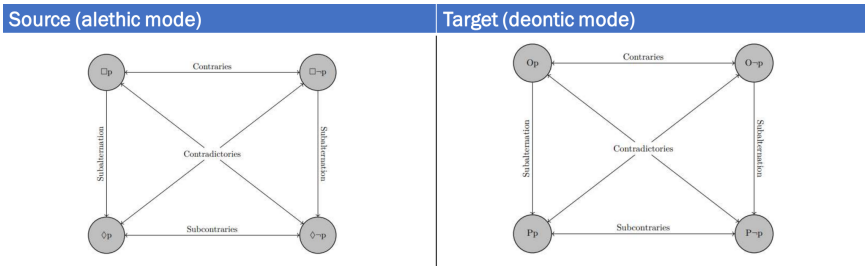
5.1.3 Analogy 3

There are **5 statuses** that can be represented by a **threefold partition**.



5.1.4 Analogy 4

There is **1 square of opposition**.



- Two propositions are **contradictories** if and only if the **truth** of one implies the **falsity** of the other.
- Two propositions are **contraries** if and only if they cannot both be **true** but can both be **false**.
- Two propositions are **subcontraries if and only if they cannot both be false** but can both be **true**.
- Two propositions are in a **subalternation** relation if and only if the **truth** of the first proposition **(superaltern)** implies the **truth** of the second **(subaltern)** but NOT vice versa.

5.1.5 In summary

Analogy	Description
Analogy 1	There are **2 operators**. Source (alethic mode): $\Box, \Diamond$ Target (deontic mode): O, P
Analogy 2	There are **5 statuses**. Source (alethic mode): **necessity, possibility, impossibility, non-necessity, contingency** Target (deontic mode): **obligatoriness, permissibility, impermissibility, omissibility, optionality**
Analogy 3	These **5 statuses** can be represented by a **threefold partition**. Source (alethic mode): **necessity, contingency, impossibility** Target (deontic mode): **obligatoriness, optionality, impermissibility**
Analogy 4	There is **1 square of opposition**. Source: (alethic mode): **modal square of opposition** Target: (deontic mode): **deontic square of opposition**

5.2 Disanalogies

Disanalogy	Source (alethic mode)	Target (deontic mode)
Disanalogy 1	$\Box p \rightarrow p$ If p is **true** across ALL possible worlds that are **accessible**, the p must be **true** in the **actual world** @.	$\neg(Op \rightarrow p)$ It does not follow from the fact that the action described by p is **obligatory** that the action is performed in the **actual world** @.
Disanalogy 2	$p \rightarrow \Diamond p$ If p is **true** in the **actual world** @, then p must be **true** in at least one possible world that is **accessible**.	$\neg(p \rightarrow Pp)$ It does NOT follow from the fact that the action is performed in the **actual world** @ that is **permissible**.

5.3 Components

Component of the logical system	Elaboration
Alphabet	The **alphabet** of **deontic logic** is an extension of the **alphabet** of **propositional logic** to include the **deontic operators** O and P.
Syntax	The **syntax** of **deontic logic** is an extension of the **syntax** of **propositional logic** to handle **well-formed formulae (wffs)** containing at least one **deontic operator**.
Semantics	The **semantics** of **deontic logic** is of the form $\langle W, S, @ \rangle$, where W denotes a **set of worlds**, S denotes a **binary relation of moral satisfaction** between worlds, and @ denotes the **actual world (privileged)**.
Proof theory	The **proof theory** of **deontic logic** comprises a set of **definitions, axioms**, and **rules of inference**. This **proof theory**, with its reliance on **axioms**, is known as **Hilbert-style proof theory**.

5.4 Proof theory

Definitions	Axioms	Rules of inference
Definition 1:	(T) All **tautologous well-formed formulae** from **propositional logic**.	($\rightarrow_{E1}$ or *modus ponens*) $p \rightarrow q$, $p \vdash q$
$Op \stackrel{\text{def}}{=} \neg P \neg p$	(NC) $\neg(Op \wedge O\neg p)$ It cannot be the case that both p and $\neg p$ are **obligatory**.	
Definition 2:	(K) $O(p \rightarrow q) \rightarrow (Op \rightarrow Oq)$ If performing the action described in p **commits** me to performing the action described in q, if p is **obligatory**, q will be **obligatory** too.	
$Pp \stackrel{\text{def}}{=} \neg O \neg P$	(NEC) $\vdash p \rightarrow \vdash Op$ If p is **tautological**, then Op is also **tautological**.	

5.5 Monotonicity of entailment (RM)

The monotonicity of entailment just means that if a sentence follows deductively from a given set of sentences, then it also follows deductively from any superset of those sentences. It is formalised as such:

$$\text{(RM): } (\vdash p \to p) \to (\vdash Op \to Oq)$$

5.5.1 Proof

1. Assume that $\vdash p \to q$
2. $\therefore (\vdash p \to q) \to \vdash O(p \to q)$
3. $\therefore \vdash O(p \to q)$
4. $\therefore \vdash O(p \to q) \to \vdash Op \to Oq$
5. $\therefore \vdash Op \to Oq$
6. $\therefore (\vdash p \to q) \to (\vdash Op \to Oq)$

5.6 Corollary of the monotonicity of entailment (Corollary of RM)

The corollary of the monotonicity of entailment is that if a given argument is deductively valid, it cannot become invalid by the addition of extra premises. It is formalised as such:

$$\text{(Corollary of RM): } \vdash Op \to O(p \vee q)$$

5.6.1 Proof

1. $\vdash p \to (p \vee q)$
2. $\therefore (\vdash p \to (p \vee q)) \to (\vdash Op \to O(p \vee q))$
3. $\therefore \vdash Op \to O(p \vee q)$

5.7 Theorem T1

$$(T1) \vdash O(p \wedge q) \to Oq$$

5.7.1 Proof

1. $\vdash (p \wedge q) \rightarrow q$
2. $\therefore (\vdash (p \wedge q) \rightarrow q) \rightarrow (\vdash O(p \wedge q) \rightarrow Oq)$
3. $\therefore \vdash O(p \wedge q) \rightarrow Oq$

5.8 Paradoxes of obligation

5.8.1 Paradox of the gentle murderer

Propositions (P):

1. It ought to be the case that A does not kill his mother.
2. If A does kill his mother, then it ought to be the case that A kills her gently.
3. A does kill his mother.

Proof:

1. $O\neg p$, where p denotes "A kills his mother" (from P1).
2. $p \rightarrow Oq$, where q denotes "A kills his mother gently" (from P2).
3. p (from P3).
4. $\therefore Oq$ (from 2, 3, and $\rightarrow_{E1}$ or *modus ponens*).
5. $\therefore \vdash q \rightarrow p$ (from T).
6. $\therefore (\vdash q \rightarrow p) \rightarrow (\vdash Oq \rightarrow Op)$ (from RM, uniformly substitute p for q and vice versa).
7. $\therefore \vdash Oq \rightarrow Op$ (from 5, 6, and $\rightarrow_{E1}$ or *modus ponens*).
8. $\therefore Op$ (from 4, 6, and $\rightarrow_{E1}$ or *modus ponens*).

Using the monotonicity of entailment, it follows that A should kill his mother, which is an odd thing to say.

5.8.2 Ross' paradox

Propositions (P):

1. It is obligatory that the letter is mailed.
2. Therefore, it is obligatory that the letter is mailed, or the letter is burnt.

Proof:

1. Op, where p denotes "the letter is mailed"
2. $\vdash Op \rightarrow O(p \vee q)$ (Corollary of RM)
3. $\therefore O(p \vee q)$, where q denotes "the letter is burnt" (from 1, 2, and $\rightarrow_{E1}$ or *modus ponens*)

It is odd to say that P1 and the corollary of RM entail an obligation that can be fulfilled by burning the letter (presumably an **impermissible** action).

5.8.3 The Good Samaritan paradox

Propositions (P):

1. It ought to be the case that A helps B, who has been robbed.
2. Therefore, it ought to be the case that B has been robbed.

Proof:

1. $O(p \wedge q$, where p denotes "A helps B" and q denotes "B has been robbed" (from P1).
2. $\vdash O(p \wedge q) \rightarrow Oq$ (T1)
3. $\therefore Oq$ (from 1, 2, and $\rightarrow_{E1}$ or *modus ponens*)

It is odd to say that from P1 and T1, it follows that B's being robbed is also obligatory.

5.9 Resolving the paradoxes

Response	Paradox of the gentle murderer	Ross' paradox	The good Samaritan paradox
Response 1: Distinguish between **non-derivatively obligatory** and **derivatively obligatory** actions.	Refrain from killing your mother (**non-derivatively obligatory**) Kill your mother (derived from RM and **impermissible**)	Mail the letter (**non-derivatively obligatory**) Burn the letter (derived from the **corollary** of RM and **impermissible**)	Help someone who is in need (**non-derivatively obligatory**) Rob the individual who has been robbed (derived from T1 and **impermissible**)
Response 2: Reject RM	RM gives rise to the **gentle murderer paradox**.	The **corollary of RM** gives rise to **Ross' paradox**.	T1, a theorem derived from RM, gives rise to the **good Samaritan paradox**.
Response 3: Introduce a **dyadic (2-placed) version of deontic logic**.	O($\neg$ murder | T) (**unconditional obligation**) O (gentle murder | murder) (**conditional obligation** if the **unconditional obligation** is violated) We cannot derive an **unconditional obligation** to murder. $\nvdash O$(murder | T)	O(mail | text has been written) (**conditional obligation**) We cannot derive an **obligation** to mail or burn the letter. $\nvdash O$(mail $\vee$ burn)	O(help B | B has been robbed) (**conditional obligation**) We cannot derive an **obligation** for B to have been robbed. $\nvdash O$(B has been robbed)

5.10 Mimamsa deontic logic

- Classical deontic logic is a **monadic (1-placed)** system:

$$O(\_)$$

- At least some systems of **deontic logic** are **dyadic (2-placed)** systems:

$$O(\_|\_)$$

It ought to be the case that A helps B and B has been robbed. Denoting p as "A helps B" and q as "B has been robbed":

- Formal representation under **monadic deontic logic**:

$$O(p \wedge q)$$

- Formal representation under **dyadic deontic logic**:

$$O(p|q)$$

5.10.1 Dyadic deontic operator $O(\phi|\theta)$

The **dyadic deontic operator** $O(\phi|\theta)$ is used in **dyadic deontic logic** to represent **conditional obligations**. ϕ represents the **main argument** and θ represents the **triggering condition**.

- It is necessarily the case that given p, it is obligatory that q.

$$\Box O(q|p)$$

- There is a **conditional obligation** that q, given p.

$$\Box O(q|p)$$

- There is an **unconditional obligation** that q, given that anything is the case.

$$O(q|T)$$

6 Inductive reasoning

$$\phi \vdash_P^I \psi$$

Where:

- ϕ is the set of premises, which potentially includes the knowledge base
- $\vdash$ represents that ψ is provable from ϕ
- I refers to the inductive mode
- P refers to the inferential mechanism of the calculus of probability
- ψ is the set of conclusions derived from the set of premises

6.1 Newcomb's paradox (Newcomb's problem)

There is a reliable predictor, another player, and two boxes designated A and B. The player is given a choice between taking only box B or taking both boxes A and B. The player knows the following:

- Box A is transparent and always contains a visible $1000.
- Box B is opaque, and its content has already been set by the predictor:
  - If the predictor has predicted that the player will take boxes A and B, then box B contains nothing.
  - If the predictor has predicted that the player will take only box B, then box B contains $1,000,000.

The player does not know what the predictor predicted, or what box B contains while making the choice.

6.2 Philosophical principles

Principle	Description
The **principle of multiple explanations**. (Epicurus, c. 300 B.C.E.)	If multiple theories H_1, H_2, and so on, are **consistent** with our observation E, then we should retain ALL these theories H_1, H_2, and son on.
The **uniformity of nature principle**. (Hume, 1739/40)	Nature is **sufficiently uniform** that **unobserved instances** in the **future** will resemble **observed instances** in the **future** will resemble **observed instances** in the **past**.
Occam's razor principle. (William of Ockham, 14th century C.E.)	Entities should NOT be multiplied beyond necessity.

6.3 Bayes' theorem

$$(BT)P(H|E) = \frac{P(E|H) \times P(H)}{P(E)}$$

Where:

- BT refers to the finite set of rules of inferences, which is **Bayes' rule** or **Bayes' theorem**.
- P is the probability of something
- H is the hypothesis
- E is the evidence
- $P(H|E)$ means the likelihood of H given E, it also refers to the **posterior probability**
- $P(E|H)$ means the likelihood of E given H
- $P(H)$ refers to the **prior probability** of hypothesis H without ANY given conditions

6.4 Conditional probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ if } P(B) \neq 0$$

Where:

- P is the probability of something
- A is an event
- B is another event
- $\cap$ is the intersection of event A and B, i.e. the probability of event A and event B happening

6.5 Deductive vs inductive reasoning

Deductive reasoning	Inductive reasoning
We reason under **certainty** concerning propositions that are either **true** or **false**.	We reason under **uncertainty** concerning propositions in which we have **differing degrees of belief**.
Deductive reasoning is **monotonic**. If $\phi \vdash \psi$, then adding more information λ to the **premise set** ϕ will NOT invalidate out **conclusion** that ψ.	**Inductive reasoning** is **non-monotonic**. Although it may be **true** that $\phi \vdash \psi$, it need NOT be the case that $(\phi \wedge \lambda) \vdash \psi$. λ may constitute **new evidence**, forcing us to **retract or revise** our **conclusion** that ψ.
Deductive reasoning is **non-ampliative**. **Deductive reasoning** unpacks the information content of the **premise set** ϕ, such that the information contained in the **conclusion set** ψ is already present (albeit in implicit form) in ϕ.	**Inductive reasoning** is **ampliative**. The information in the **conclusion** that ψ **exceeds and amplifies** the information content of the **premise set** ϕ.

6.6 Axioms of probability (Kolmogorov theorem)

6.6.1 Degrees of belief

The **degrees of belief** are constrained by a finite set of **axioms of probability**. Any probability function P must satisfy the following axioms:

Axiom	Description
K1 (**non-negativity**)	$P(A) \geq 0$ in **sample space** Ω, where $P(A)$ is the **probability** of **outcome** A.
K2 (**normalisation**)	$P(\Omega) = 1$
K3 (**addition rule**)	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ If A and B are **mutually exclusive**, then $(A \cap B) = \emptyset$ and $P(A \cap B) = 0$. $\therefore P(A \cup B) = P(A) + P(B)$
K4 (**complement rule**)	$P(\bar{A}) = P(\Omega) - P(A) = 1 - P(A)$

6.6.2 Ruled-out scenarios

A set of **outcomes** is **jointly exhaustive** if these **outcomes** encompass the entire **sample space** Ω. In other words, at least one of these **outcomes** must occur.

The **axioms of probability** rule out the following scenarios:

Scenario	Axiom ruling out the scenario
The assignment of **negative probability values** to individual **outcomes**	K1 or **non-negativity**
The assignment of **probability values** to **jointly exhaustive and mutually exclusive outcomes** that sum to > 1.	K2 or **normalisation**
The assignment of **probability values** to **jointly exhaustive outcomes** that sum to < 1.	K2 or **normalisation**
The assignment of a **probability value** other than $1 - p$ to $\bar{A}$, when an agent assigns a **probability value** p to some **outcome** A.	K4 or **complement rule**

6.7 Axioms of expected utility theory (von-Neumann-Morgenstern theorem)

6.7.1 Degrees of preference

The **degrees of preference** are constrained by the **axioms of expected utility theory**.
Any agent faced with a system U of alternative entities $u, v, \ldots$ must satisfy the following axioms:

Scenario	Axiom ruling out the scenario
VM1 (**completeness**)	For every u and v, one and only one of the following relations holds: $u \succ v$ (the agent **prefers** u to v) $v \succ u$ (the agent **prefers** v to u) $u \sim v$ (the agent is **indifferent** between u and v) Alternatively, for every u and v, either $u \succeq v$ or $v \succeq u$.
VM2 (**transitivity**)	For every u, v and $w, u \succ v$ and $v \succ w$ imply that $u \succ w$. Alternatively, for every u, v, and w, if $u \succeq v$ and $v \succeq w$, then $u \succeq w$.
VM3 (**independence of irrelevant alternatives**)	For every u, v, and w, suppose that $u \succeq v$ and a **third irrelevant alternative** w is present. The **order of preference** of u over $v (u \succeq v)$ holds, independently of the presence of absence of the **third irrelevant alternative** w.
VM4 (**continuity**)	Let L denote a **lottery** whose 2 possible **outcomes** are u and v, $L \stackrel{\text{def}}{=} \{u, v\}$ $P(u) = \alpha$ and $P(v) = 1 - \alpha)$, where $0 < \alpha < 1$ For every u, v, and $w, v \succ w \succ u$ implies the following: The existence of an α such that $w \succ L$ when $1-\alpha$ or $P(v)$ is **sufficiently small**. The existence of an α such that $w \sim L$ at a certain value of $(1 - \alpha)$. The existence of an α such that $L \succ w$ when $1-\alpha$ or $P(v)$ is **sufficiently large**.

6.7.2 Ruled-out scenarios

The **axioms of expected utility theory** rule out the following scenarios:

Scenario	Axiom ruling out the scenario
The agent prefers neither alternative to another nor remains **indifferent** between both alternatives.	VM1 (**completeness**)
The agent preferring u to v and v to w but remaining **indifferent** between u and w.	VM (**transitivity**)
The **decoy effect**.	VM3 (**independence of irrelevant alternatives**)
The impossibility of an agent preferring lottery L to w, where $v \succ w \succ u$ and $L \stackrel{\text{def}}{=} \{u, v\}$	VM4 (**continuity**)

- According to the **Cox-Jaynes model**, any system **reasoning under uncertainty** and in terms of **degrees of belief** will conform to the **axioms of probability**.
- Furthermore, if the **axioms of expected utility theory** are satisfied, then the agent is said to be **rational** and the **preferences** can be represented by a **utility function**.

6.8 Bayesian decision theory

- **Standard decision theory** addresses **individual decision-making** under **uncertainty**.
- **Standard decision theory** incorporates the **axioms of probability** (K1 - K4) and the **axioms of expected utility theory** (VM1 - VM4).

Bayesian decision theory incorporates **standard decision theory** and **Bayesian epistemology** (BT):

Step	Description
1	Identify n alternative courses of action $\phi_1, \phi_2, \ldots, \phi_n$ and their m associated possible outcomes, where $\{M, n\} \in \mathbb{N}$.
2	Characterise each action ϕ_i in terms of its possible outcomes s_j, where $\{i, j\} \in \mathbb{N}, i \in (0, n]$ and $j \in (0, m]$.
3	Assign **prior probabilities** to each outcome $P(s_j \vert \phi_i)$ in accordance with K1 - K4.
4	Assign **utility values** to each outcome $U(s_j \cap \phi_i)$ in accordance with VM1 - VM4.
5	Gather evidence and **update probabilities** by applying Bayesian epistemology.
6	Multiplying the **updated probability** and the **utility values** each **outcome** s_j relative to ϕ_i.
7	Sum the products across each section ϕ_i to determine its **expected utility**.
8	Select the action ϕ_i with the **highest expected utility** as the **morally right action**.

6.9 Jeffrey-Bolker theory

Jeffrey-Bolker's expected utility theory is an example of **evidential decision theory**. It relies on a boolean algebra Ω that consists of:

Formal representation	Description
A, B, C, etc.	**Propositions** as elements of Ω.
$\bar{A}$	**Negation**, such that if $A \in \Omega$, then $\bar{A} \in \Omega$.
$A \cup B$	**Disjunction**, such that if $A, B \in \Omega$, then $A \cup B \in \Omega$.
$\top$	**Tautology**.
$\bot$	**Contradiction** or negation of $\top$.
$\succeq$	A coherent **preference order relation** over Ω'.

6.9.1 Strategy

Jeffrey aims to recommend a **Bayesian model of deliberation** and a corresponding **theory of preference**.

Move	Description
1	Identify the **Bayesian principle of deliberation**. According to this principle, we rank actions $\phi_1, \phi_2, \ldots, \phi_n$ in order of **preference**. Given a **coherent preference ranking**, we can discover a pair of **probability and desirability assignments** (roughly corresponding to the **probability** and **utility value assignments**) to propositions describing the performance of these actions.
2	Introduce the **coherence** assumption. According to this assumption, the agent's **preference ranking** has the following properties: Property 1: **Coherence** There is an underlying set of **probabilities and desirabilities** that yield the **preference ranking** via the **Bayesian principle of deliberation**. Property 2: **Defeasibility** **Experience and reflection** constantly force the agent to **revise their agent preference ranking**.
3	Characterise the **desirability function (des)** and the **probability measure (prob)**. The **desirabilities (des)** of the **basic cases** may be any set of numbers, independent of the **probabilities** des$A > 0$ if A is **good**, des $A = 0$ if A is **indifferent**, des $A < 0$ if A is **bad**. The **probabilities (prob)** of the **basic cases** may be any set of non-negative numbers that sum to 1 ($P(A) \geq 0, P(\Omega) = 1$).

6.9.2 Example

Event	L (live to age 65 or more)	$\bar{L}$ (die before age 65)
S (smoke)	Best ($S \cap L$)	3^{rd} best ($S \cap \bar{L}$)
$\bar{S}$ (quit)	2^{nd} best ($\bar{S} \cap L$)	Worst ($\bar{S} \cap \bar{L}$)

6.9.3 Non-Bayesian deliberation of the example

According to the **syllogistic line of reasoning**:

- P1: Either L or $\bar{L}$.
- P2: If L, then S is more desirable than $\bar{S}$.
- P3: If $\bar{L}$, then (equally) S is more desirable than $\bar{S}$.
- C: $\therefore S$ is more desirable than $\bar{S}$ (**fallacious**)

This **fallacious line of reasoning** wrongly assumes that the **4 possible action-outcome pairs** are **equiprobable**:

$$P(S \cap L) = P(S \cap \bar{L}) = P(\bar{S} \cap L) = P(\bar{S} \cap \bar{L}) = 0.25$$

$$\begin{aligned}
U(S) &= \frac{P(S \cap L) \times v(S \cap L) + P(S \cap \bar{L}) \times v(S \cap \bar{L})}{P(S \cap L + P(S \cap \bar{L}))} \\
&= \frac{P(S \cap L) \times v(S \cap L) + P(S \cap \bar{L}) \times v(S \cap \bar{L})}{P(S)} \\
&= P(L|S) \times v(S \cap L) + P(\bar{L}|S) \times v(S \cap \bar{L}) \\
&= \frac{0.3}{0.5} \cdot 100 + \frac{0.2}{0.5} \cdot -90 \\
&= 60 - 36 \\
&= 24
\end{aligned}$$

6.9.4 Bayesian deliberation of the example

Suppose that a **probability measure** P allows us to assign the following **probability values**:

$$P(\Omega) = P(S) + P(\bar{S}) = 1$$

$$P(S) = P(\bar{S}) = 0.5$$

$$P(S) = P(S \cap L) + P(S \cap \bar{L})$$

$$P(\bar{S}) = P(\bar{S} \cap L) + P(\bar{S} \cap \bar{L})$$

$$P(S \cap L) = 0.3$$

$$P(S \cap \bar{L}) = 0.2$$

$$P(\bar{S} \cap L) = 0.4$$

$$P(\bar{S} \cap \bar{L}) = 0.1$$

Suppose a **desirability measure** v allows us to assign the following **desirability values**:

$$v(S \cap L) = 100 \text{ (Best)}$$

$$v(\bar{S} \cap L) = 70 \text{ (2nd best)}$$

$$v(S \cap \bar{L}) = -90 \text{ (3rd best)}$$

$$v(\bar{S} \cap \bar{L}) = -100 \text{ (Worst)}$$

$$\begin{aligned} U(\bar{S}) &= P(L|\bar{S}) \times v(\bar{S} \cap L) + P(\bar{L}|\bar{S}) \times v(\bar{S} \cap \bar{L}) \\ &= \frac{0.4}{0.5} \cdot 70 + \frac{0.1}{0.5} \cdot -100 \\ &= 56 - 20 \\ &= 36 \end{aligned}$$

$$\therefore U(\bar{S}) > U(S)$$

$\therefore \bar{S} \succ S$ (**quitting** is preferable to **continuing to smoke**)

7 Issues with standard decision theory

7.1 Consistency versus responsibility

- The **axioms of probability** (K1 - K4) and the **axioms of expected utility theory** (VM1 - VM4) provide important constraints on the assignment of **probability and utility values**.
- However, consistency with these axioms is insufficient to ensure **responsible decision-making**.
- For example, think about an individual who is a **flat earth theorist** who thinks that the **flat earth theory** is 100% correct and all other theories are wrong. Such an individual is **Kolmogorov-consistent**, as his beliefs conform to the **axioms of probability**, but is **epistemically irresponsible**.
- Another example would be someone who prefers **genocide to murder** and **murder to a walk in the hills**. Such a person is **von-Neumann-Morgenstern-consistent** as his **degrees of preference** conform to the **axioms of expected utility theory**, but is **morally irresponsible**.

7.2 Cognitive biases

- Humans are not perfect, and hence we all suffer from cognitive biases.
- One example is the **decoy effect**, which is a **cognitive bias** in which consumers demonstrate a **shift in preferences** between two options when presented with a **third option that is asymmetrically dominated**.
- An example of this effect would be having 3 options for a product, but 1 option is worse than all other options, such as the following:

Option	Description	Price
Option 1	Online subscription for a newspaper	59.00
Option 2	Print subscription for a newspaper	125.00
Option 3	Print and online subscription for a newspaper	125.00

- When option 2 is removed, some people may change their preference from option 3 to option 1, which violates the **independence of irrelevant alternatives**.

7.3 Deontological decision-making

- There are difficulties in modelling **deontological decision-making**.
- We denote the utility value of a **prohibited or impermissible** outcome as $-\infty$ and the utility value of an **obligatory** outcome as $+\infty$.
- This sets **prohibitions** and **obligations** apart from other **actions**, as their associated **outcomes** have **absolute maximum or minimum expected utility**.

Problem	Description
Problem 1: Swamping out of **probability values**.	An **infinite utility or disutility value** completely **swamps out any probability value** associated with an outcome. If **killing** is **morally prohibited**, then the outcome of **murder** will be assigned the utility value of $-\infty$, hence, any actions that may lead to murder, no matter how unlikely, will presumably also be **morally prohibited**.
Problem 2: All **prohibited** or **obligatory** actions are on par	**Murder** is no better or worse than **genocide**.
Problem 3: Violation of the **continuity** axiom.	Any preservation of the continuity axiom could be questioned on **deontological grounds**. Suppose that w is **forbidden** and u and v are **permissible**, such that $v \succ u \succ w$. $L \stackrel{\text{def}}{=} \{v, w\}$ According to the continuity axiom, there will be a probability such that the permissible action u is as good as a lottery involving another **permissible and more preferable action** v and a **prohibited action** w. However, according to **deontology**, any lottery involving a **prohibited action**, no matter how unlikely, cannot be ranked as on par with an ordinary **permissible action**.

7.4 Collective decision-making and voting paradoxes

Using the preference data below:

Individual	Preference
1	$A \succ C \succ D \succ B$
2	$B \succ C \succ D \succ A$
3	$D \succ A \succ C \succ B$
4	$A \succ B \succ D \succ C$
5	$D \succ A \succ C \succ B$

7.4.1 Condorcet method

The Condorcet method conducts **pairwise comparisons**, and the winner is the choice that wins all the head-to-head matchups.

Head-to-head match-up	Winner	Score
A versus B	A	4-1
A versus C	A	4-1
A versus D	D	2-3
B versus C	C	2-3
B versus D	D	2-3
C versus D	D	2-3

7.4.2 Borda count

The Borda count method assigns points to the preference order of the choices. In this example, 0 points are awarded to the last choice, 1 point to the second-last choice, 2 points to the second choice, and 3 points to the first choice.

Choice	Number of points	Score
A	$3 + 0 + 2 + 3 + 2$	10 pts
B	$0 + 3 + 0 + 2 + 0$	5 pts
C	$2 + 2 + 1 + 0 + 1$	6 pts
D	$1 + 1 + 3 + 1 + 3$	9 pts

7.4.3 Problems

Problem	Description
Problem 1: Different methods yield different winners	D is the **Condorcet winner**, whereas A is the winner under the **Borda count** method.
Problem 2: **Voting paradoxes**	**Individual preferences** may be **non-cyclic** and consistent with the axioms VM1 - VM4. However, **collective preferences** could be **cyclic**. This is known as the **Condorcet paradox** or the **voting paradox**. Individual 1: $A \succ B \succ C$ Individual 2: $B \succ C \succ A$ Individual 3: $C \succ A \succ B$ Collectively: $A \succ B \succ C \succ A$ (**paradoxical**) While **individual preferences** may obey VM1 (**completeness**) and VM2 (**transitivity**), it by no means follows that **collective preferences** will obey VM1 and VM2. P(**Condorcet paradox**) $\approx 9\%$ (low)

7.4.4 Arrow's impossibility theorem

When voters have more than 3 distinct choices, **no ranked voting electoral system** can **convert the ranked individual preferences into ranked collective preferences** while satisfying the following conditions:

Condition	Description
1	**Pareto efficiency**
2	**Independence of irrelevant alternatives** (VM3)
3	**Unrestricted domain**
4	**Absence of dictatorship**

Alternatively, there is no constitution by which **ranked individual preferences** can be aggregated into **ranked collective preferences**, while satisfying **basic fairness criteria**, unless there is a **dictatorship** (not condition 4).

8 Game theory

8.1 Comparison with decision theory

Decision theory	Game theory
The concern is with **individual decision-making**.	The concern is with **interdependent decision-making**.
An individual's choice is neither affected by nor affecting the choices of other individuals.	An individual's choice affects the choices of other individuals. Each individual also has to consider the **utility functions** of other individuals and how they will affect the choices of other individuals and the overall outcome.

8.2 Prisoner's dilemma

- There are two suspects, P1 and P2.
- The district attorney believes that P1 and P2 are **guilty of a crime** but does not have sufficient evidence to convict them.

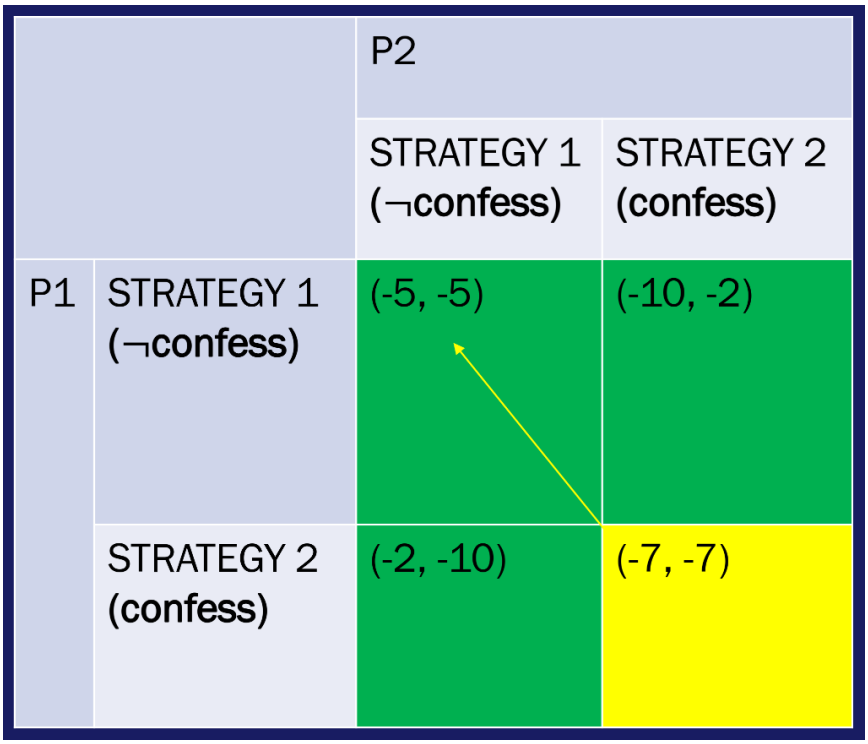
Each of P1 and P2 has 2 strategies:

Strategy	Description
1	Do not confess the crime (cooperate)
2	Confess the crime to the police (defect)

P1 and P2 are confronted with 4 options relative to various **possible strategy combinations**:

Option 4: If both P1 and P2 **do not confess**, then they will each get **1 year in prison**.	Option 3: If P2 **confesses** and P1 does not, P2 will get **3 months in prison** and P1 will get **10 years in prison**.
Option 2: If P1 **confesses** and P2 does not, P1 will get **3 months in prison** and P2 will get **10 years in prison**.	Options 1: If both P1 and P2 **confess**, then they will each get **8 years in prison**.

8.3 Table of utility values



- The cells coloured in **green** denotes a **Pareto-optimal** state of affairs.
- The cell coloured in **yellow** denotes the **Nash equilibrium**.
- If P1 and P2 are allowed to communicate and bargaining is **cost-free**, then P1 and P2 could agree to **cooperate** and not **confess**.
- Hence, they could make a **Pareto-efficient move** indicated by the arrow ($\rightarrow$)
- The **Nash equilibrium** (coloured in **yellow**) arises because P1 and P2 behave as **straightforward maximisers**.
- However, P1 and P2 have reasons to become **constrained maximisers**.

8.4 Modification of rationality assumption

- The original **rationality** assumption, which is **straightforward maximisation**, is as such: It is **rational** to choose the course of action with the **maximum expected utility**.
- The modified **rationality assumption**, is as such: It is **rational** to be disposed to **constrained maximisation** and **cooperate** in **prisoner's dilemma-type scenarios**.

8.5 Possible strategies for the game

- Random, which is to choose to cooperate 50% of the time.
- Tit-for-tat (TFT), which is to choose to cooperate on the first move, and choose your opponent's last move as your next move.
- Suspicious tit-for-tat (STFT), which is to choose to defect on the first move, and choose your opponent's last move as your next move.
- Tit for two tats (TF2T), which is to choose to cooperate on the first two moves, then choose to cooperate as the next move, unless your opponent chooses to defect for 2 moves. When your opponent stops choosing to defect, then choose to cooperate.

8.6 Axelrod's tournaments

- The prisoner's dilemma was originally introduced as a **2-player game**, but it was later embedded by Axelrod into **round-robin tournaments**.
- Each program was pitted against the rest of the field.
- The aim of these tournaments was to learn about how to **choose effectively** in an **iterated prisoner's dilemma**.

8.6.1 Properties of successful strategies

Property	Description
Be **nice**	Choose to cooperate on the first move. For example, Cooperative, TFT, TF2T.
Be **forgiving**	Do not immediately retaliate if your opponent chooses to defect in a move. For example, Cooperative, TF2T.
Be **prepared to retaliate if necessary**.	You must be prepared to choose to defect at some point if your opponent keeps choosing to defect. For example, TFT, STFT, TF2T

- What accounts for **tit-for-tat (TFT)'s** success is its combination of being **nice, retaliatory, forgiving,** and **clear**.
- Its **niceness** prevents it from getting into **unnecessary trouble**.
- Its **retaliation** discourages the other side from persisting whenever **defection** is tried.
- Its **forgiveness** helps restore **mutual cooperation**.
- Its **clarity** makes it intelligible to the other player, thereby eliciting **long-term cooperation**.

8.7 Rapoport et al.'s objections to the Axelrod tournaments

8.7.1 Objection 1

The choice of **tournament format**:

- In a **knockout tournament**, top-ranked contestants at each stage progress to the next stage.
- As the tournament continues, the number of competitors decreases.
- In a **round-robin tournament**, each contestant competes with each of the others an equal number of times.
- Axelrod chose the **single-stage round-robin format** for his tournaments.
- He provided no **justification** for this choice of **tournament format**.
- The **single-stage round-robin format** becomes **impractical** when the **number of contestants is large**, although this problem disappears when the tournament is run on a computer.

8.7.2 Objection 2

The choice of **criterion for determining success**:

- The **criterion for determining success** involved **maximising the number of points won across all dyadic interactions**.
- Axelrod chose this **criterion of success**, but once again provided no **justification** for this choice of criterion for determining success.
- Most of the programs were not designed to **maximise the total number of points**.
- **Tit-for-tat (TFT)** can never win any particular **iterated prisoner's dilemma game**.
- **Tit-for-tat (TFT)** can never achieve a **positive point difference** against any other program.

8.7.3 Objection 3

The choice of **payoff structure**:

- The 2×2 **prisoner's dilemma payoff matrix** had conventional values, where T denotes **sole defection**, R denotes **joint cooperation**, P denotes **joint defection**, and S denotes **sole cooperation**.
- The values in this matrix are $(T, R, P, S) = (5, 3, 1, 0)$.
- Axelrod chose this **payoff structure** but again provided no **justification** for this choice of **payoff structure**.

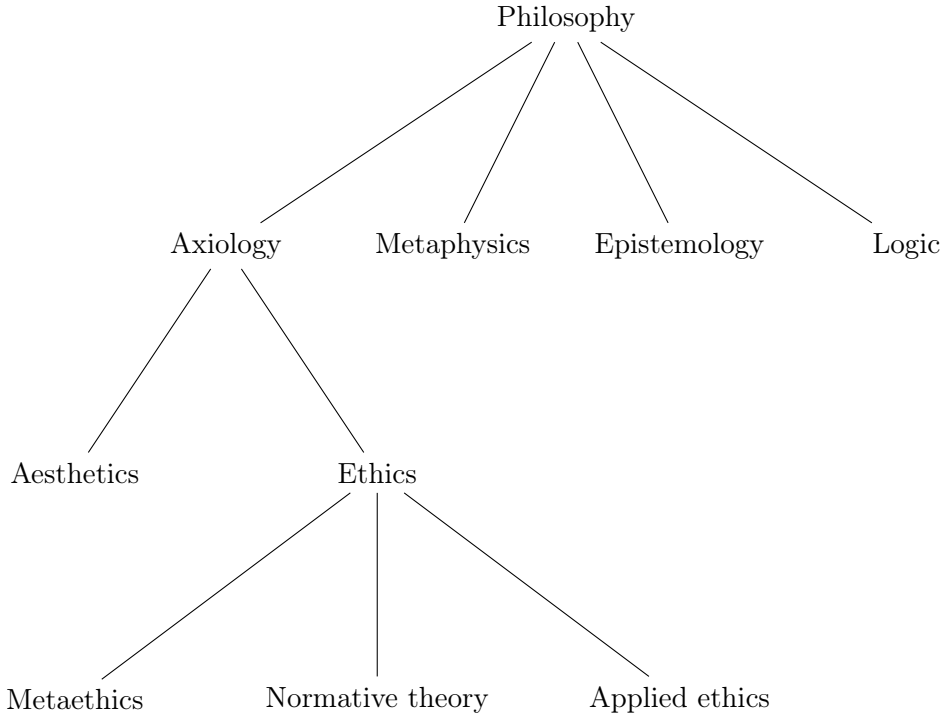
8.7.4 Conclusion

- Once the **2-player prisoner's dilemma** is embedded into a **tournament**, decisions have to be made about the **tournament format** (objection 1), **criteria for determining a winner** (objection 2), and **payoff structure** (objection 3).
- However, Axelrod has provided no **justification** for his choices of **tournament format, criterion of determining success**, and **payoff structure**.
- Hence, the policy recommendations about the effectiveness of **tit-for-tat (TFT)** should be qualified, i.e. the recommendations cannot be generalised.

9 Ethics

Ethics and **moral philosophy** are children nodes of the branch of **axiology**. Children nodes of **ethics** and **moral philosophy** include:

- **Metaethics, normative theory**, and **applied ethics**.



9.1 Branches of ethics

Branch of ethics and moral philosophy	Central question	Elaboration
Metaethics (2^{nd}-order theory)	What is **morality**?	**Metaethics** is concerned with the **status, foundation, and scope** of **moral facts, values, properties, and terms**.
Normative theory (1^{st}-order theory)	What is **moral** (in **general**)?	**Normative theory** is concerned with the articulation of **moral standards** and **rules of conduct**.
Applied ethics (praxis)	What is **moral** (in **specific, controversial issues**)?	**Applied ethics** is concerned with the application of **philosophical theory** to **practical problems**.

9.2 Normative neutrality requirement

Branch of ethics and moral philosophy	Description
Normative theory (1^{st}-order theory)	There is NO **normative neutrality requirement**. **Normative theory** must have **substantive moral commitments**.
Metaethics (2^{nd}-order theory)	There is a **normative neutrality requirement**. **Metaethics** can have **classificatory moral commitments** but must avoid **substantive moral commitments**.

9.3 Forcehimes' collapse argument

- P1: There is a requirement for **normative neutrality** in **metaethics**.
- P2: Such a **breach of normative neutrality** is **inevitable**.
- Conclusion: Hence **metaethical theories (2nd-order)** turn out to be **normative theories (1st-order)** in disguise.

9.4 Normative theory

The different approaches to **normative theory (1st-order)** give rise to different **substantive moral commitments** and different **moral standards** and **rules of conduct**. The different approaches include:

Degree of particularity	Approach	Elaboration
General	Approach 1: **High moral theory**	**Consequentialism** (camp 1), **deontology** (camp 2), **virtue ethics** (camp 3)
↓	Approach 2: **Mid-level theory**	**Autonomy** (principle 1), **beneficence** (principle 2), **non-maleficence** (principle 3), **justice** (principle 4)
↓	Approach 3: **Casuistry** or **case-based reasoning**	A **bottom-up approach** in which **moral principles** and **moral theories** emerge from case-based moral judgments.
Particular	Approach 4: **Narrative ethics**	We use stories to make sense of our experiences.

9.5 High moral theory

From a set of n alternative courses of action, where $i, n \in \mathbb{N}$ and $i \in (0, n]$:

Camp	Description
Consequentialism	ϕ_i-ing is **morally right** if and only if it **maximises the good**, where the **good** is defined in terms of **some theory of the good T**.
Deontology	ϕ_i-ing is **morally right** if and only if it has **intrinsic moral worth**.
Virtue ethics	ϕ_i-ing is **morally right** if and only if it is the **best action** (in terms of **virtues and vices**) that a **virtuous agent** might perform in the circumstances.

9.5.1 Advantages

Advantage	Elaboration	Substantiation
Advantage 1	**High moral theory** can provide **structured and systematic moral guidance**.	Camps 1 to 3 provide us with **moral standards** and **rules of conduct** for identifying the **morally appropriate action** (ϕ_i-ing) from n alternative courses of action.
Advantage 2	**High moral theory** can yield **moral standards** that can yield **moral standards** that can help us to achieve **consistency** and **coherence** in our moral lives.	For camp 1, there is the transitivity rule, which states if ϕ_1 is better than ϕ_2 and ϕ_2 is better than ϕ_3, then ϕ_1 must be better than ϕ_3. For camp 2, there is the no-contradiction rule, which states that one and the same action ϕ_i cannot be both **obligatory** and **impermissible**. For camp 3, there is the doctrine of the mean, which states that the **virtues** are a **mean** between the **vices of defect** and **excess**.
Advantage 3	**High moral theory** often has the relevant tools and resources for **moral justification**.	Camp 1 delivers **evaluative claims** in terms of the **maximisation of the good**. Camp 2 delivers **deontic verdicts** in terms of **duties, rights, and obligations.** Camp 3 delivers **virtue-ethical judgments** in terms of the **language of virtues and vices**.

9.5.2 Disadvantages

1. How do we choose a **moral theory** from the **competing alternatives**?
    - Even if we do make a choice, how do we **justify** that choice?
    - Individuals with **different theoretical starting points** must still agree on a **similar set of principles**.
    - Hence, it has been argued that we should move to **mid-level theory** and a more **principle-based approach**.
2. How do we navigate **disagreement** within the ranks of each camp?
    - For camp 1, **act** versus **rule based** forms of consequentialism, **maximising** versus **satisficing** forms of consequentialism.
    - For camp 2, **monistic** versus **pluralistic** forms of **deontology, agent** versus **patient-centred forms** of **deontology**.
    - For camp 3, **eudamimonist versus non-eudaimonist** forms of **virtue ethics**.
3. **High moral theory** may be too **ill-equipped** to deal with **practical decision-making at the concrete level**.
    - When the **applied ethical problems** are **complex**, how likely is that **high moral theory** will be **sufficiently fine-grained** to generate responses?
    - For a move away from a **top-down approach** and toward a greater degree of **particularity**, it has been argued that we should favour **casuistry** or **case-based reasoning** or **narrative ethics** instead.

10 Consequentialism

- From a set of n alternative courses of action, ϕ_i-ing is **morally right** if and only if it **maximises the good**, where $i, n \in \mathbb{N}, i \in (0, n]$, and the **good** is defined in terms of **some theory of the good T**.
- The **core consequentialist commitment** is **maximising the good** (however the good might ultimately be defined).

10.1 Consequentialist decision-making

1. Compare the relative merits of n alternative courses of action $\phi_1, \phi_2, \ldots, \phi_n$, where $n \in \mathbb{N}$.
2. **Evaluate** these n courses of action in terms of whether they **maximise the good**.
    - ϕ_1 **maximises the good**.
    - ϕ_2 does not **maximise the good**.
3. Arrive at the **decision outcome**. The **morally right** action is the one that **maximises the good** (for instance, ϕ_1).

10.2 Theory of the Good

- The **good to be maximised** is determined in terms of a **theory of the good T**.
- However, there are **multiple theories of the good**.
- Hence, one has not adopted any particular moral system in adopting **consequentialism** unless one says what the **good** is.

Candidate **theories of the good** include:

1. The **good** is defined as things which are **pink with yellow trimmings**.
    - This is meant to be a joke.
2. The **good** is defined as things which ought to be **maximised**.
    - This is possibly **trivial**, as the **core consequentialist commitment** becomes "that which **ought to be maximised, ought to be maximised**".
3. The **good** is defined as things that **facilitates self-interest**.
4. The **good** is defined as things that **facilitates human pleasure and happiness**.
5. The **good** is defined as that which is best understood as a **plurality of goods (happiness, justice, fairness, and so on)**.

10.3 Types of consequentialism

Distinct types of **consequentialism** can be generated from **multiple theories of the good**:

Type of consequentialism	Core consequentialist commitment	Theory of the good	Consequentialist outcome
Ethical egoism	The **good ought to be maximised**.	**Egoistic** theory of the good.	Select the action ϕ_i that **maximises the good**.
Utilitarianism		**Hedonistic** theory of the good.	
Pluralistic consequentialism and ideal utilitarianism		**Pluralistic** theory of the good.	

10.4 Types of utilitarianism

For both **act** and **rule-based utilitarianism**, the **good** is what **facilitates human pleasure, happiness,** and **utility-based considerations**.

Type of utilitarianism	Theory of the good	Description
Act-based utilitarianism	The **good** is defined as that which **facilitates human pleasure and happiness**.	ϕ_i-ing is morally right if and only if it **maximises the good**.
Rule-based utilitarianism		ϕ_i-ing is morally right if and only if it is in accordance with a certain **set of rules R** that has been selected for its **good consequences**.

10.5 Hedonism

- Both **act** and **rule-based utilitarianism** are characterised in terms of a **hedonistic** theory of the good.
- Psychological hedonism: Only **pleasure (happiness)** or **pain (unhappiness)** motivates us.
- Ethical hedonism: Only **pleasure (happiness)** has **value** and only **pain (unhappiness)** has **disvalue**.

According to the **greatest happiness principle**:

- Happiness is defined as **pleasure** and the **absence of pain** (pleasure $\wedge$ $\neg$ pain)
- Unhappiness is defined as **pain** and the **privation of pleasure** (pain $\wedge$ $\neg$ pleasure)

Types of **ethical hedonism** include:

- **Quantitative hedonism**, which states that the **quantity of pleasure (happiness)** that matters.
- **Qualitative hedonism**, which states that the **quality of pleasure (happiness)** that matters.

10.6 Benthamite utilitarianism

Benthamite utilitarianism is a **traditional account of utilitarianism** that can be characterised in terms of the following:

Core consequentialist commitment	Hedonistic theory of the good	Consequentialist outcome
The good ought to be maximised.	Psychological hedonism Ethical hedonism Quantitative hedonism	Select the action ϕ_i that **maximises the good**.

10.7 Hedonic calculus

- Bethamite utilitarianism relies on a **hedonic calculus** or **felicific calculus**.
- The **hedonic calculus** is an algorithm, formulated by Bentham, for **calculating the total quantity of pleasure (happiness)** that an action ϕ_i is likely to cause.

The **variables** in the **hedonic calculus** include:

Variable	Description
Intensity	How **strong** the **pleasure** or **pain** is.
Duration	How **long** the **pleasure** or **pain** lasts.
Probability	How **likely** the **pleasure** or **pain** is to be the result of ϕ_i-ing.
Propinquity or remoteness	How close the sensation of **pleasure** or **pain** is to be the result of ϕ_i-ing.
Fecundity	How likely ϕ_i is to lead to further **pleasures** or **pains**.
Purity	How much **intermixture** there is between **pleasure** or **pain** and other sensations.

10.8 Issues with quantitative hedonism

- Suppose that our choice is between **playing push-pin** (ϕ_1) and **reading poetry** (ϕ_2).
- Suppose that the **total net utility value** of ϕ_1 equals the **total net utility value** of ϕ_2.
- Quantitative hedonism states that only the **quantity of pleasure (happiness)** matters.
- Hence, both **Benthamite utilitarianism** and its **machine-based implementation** *Jeremy* will concur that **playing push-pin** (ϕ_1) is as good as **reading poetry** (ϕ_2).
- There is a danger that **quantitative hedonism** might lead us to overvalue **bestial, unsophisticated, lower-quality, and debauched pleasures**.
- However, human beings are able to distinguish between **lower-quality pursuits** such as ϕ_1 and **higher-quality intellectual pursuits** such as ϕ_2.

10.8.1 Possible responses

Response	Justification	Elaboration
Defend **quantitative hedonism**.	The **pleasures** associated with **reading poetry** (ϕ_2) are **more probable, more durable, more fecund or more likely to lead to further pleasures,** and **purer** (unlikely to be mixed with pain).	The **quality of pleasure** can still be reduced to **quantitative considerations**. Hence, we can still retain **quantitative hedonism**.
Ditch **quantitative hedonism** in favour of **qualitative hedonism**.	**Qualitative hedonism** gives us automatic reasons to favour **higher-quality intellectual pursuits** such as ϕ_2 over **lower-quality pursuits** such as ϕ_1.	

10.9 Qualitative hedonism

- The standard view of qualitative hedonism is that a **higher-quality pleasure** will be preferred to **any amount** of a **lower-quality pleasure**.
- Let ϕ_i denote **playing push-pin** and let ϕ_2 denote **reading poetry**.
- Suppose that the total net utility value of $\phi_1 = 5,000,000$ units and the total net utility value of $\phi_2 = 1$ unit.

  $$U(\phi_2) < U(\phi_1)$$

  However, since a higher quality pleasure is preferred to a lower quality pleasure:

  $$\phi_2(\text{higher-quality}) \succ \phi_2(\text{lower-quality})$$

- We could postulate an **infinite superiority** of **higher-quality** over **lower-quality pleasures**.
- However, this will reduce **qualitative** to **quantitative considerations**.
- For the **non-standard view** of qualitative hedonism:
  - **Qualitative hedonism** is not **quantitative hedonism** in disguise.
  - Rather, **quantity** and **quality** are two distinct properties of **pleasure**.
  - We may sometimes have to make **trade-offs** between a **lower-quantity** of a **higher-quality pleasure** and a **higher quantity** of a **lower-quality pleasure**.

10.9.1 Dilemma regarding qualitative hedonism

Argumentative constituent	Elaboration
P1: Either the **quality of pleasure** contributes to the **total net utility value** in the same manner as the other **quantitative variables** in the **hedonic calculus**, or it does not.	The other **quantitative variables** in the **hedonic calculus** are **intensity, duration, probability, propinquity** or **remoteness, fecundity**, and **purity**.
P2 (horn 1): If the **quality of pleasure** contributes to the **total net utility value** in the same manner as the other **quantitative variables**, then **qualitative hedonism** will turn out to be **quantitative hedonism** in disguise. P3 (horn 2): If the **quality of pleasure** does not contribute to the **total net utility value** in the same manner as the other **quantitative variables**, the **qualitative hedonism** will be **inconsistent**.	This is the same problem confronting the **standard view of qualitative hedonism**, when it postulates an **infinite superiority** of **lower-quality pleasure** over **higher-quality pleasure**. According to **ethical hedonism**, only **pleasure (happiness)** has **value** and only **pain (unhappiness)** has **disvalue**. Hence, **ethical hedonism** implies **value monism**. However, **quality** appears to count as another **intrinsic good**.
C: Therefore, either **qualitative hedonism** will turn out to be **quantitative hedonism** in disguise (horn 1) or **qualitative hedonism** will be **inconsistent** (horn 2).	This is the **dilemma** confronting **qualitative hedonist**.

10.9.2 Schmidt-Petri's response to the dilemma

We may sometimes have to make **trade-offs** between a **lower quantity** of a **higher-quality pleasure** and a **higher quantity** of a **lower-quality pleasure**.

1. Identify the **opponent**.
    - The standard view of **qualitative hedonism** states that a **higher-quality pleasure** will **always** be chosen over a **lower-quality pleasure**, even when the **lower-quality pleasure** is available is a **lower quantity**.
2. Identify objections to the **standard view of qualitative hedonism**.
    (a) **Ambiguity** over the notion of "**quality**". Mill does not tell us what would correspond to the concept of "**quality**".
    (b) **Lack of clarity** about **experts**. We do not know how to tell who is an **expert** in the real world.
3. Identify the **source material**, like Mill's Utilitarianism (Chapter 2, paragraph 5).
4. Distinguish between the **standard view** and the **non-standard view** relative to the **source material**.
    - The **standard view** (incorrect): $q \rightarrow p$ If we are justified in saying that x is of a **higher quality** than y, then some pleasure x is chosen over another pleasure y available in a **higher quantity**.
    - The **non-standard view** (correct): $p \rightarrow q$ If some pleasure x is chosen over another pleasure y available in a **higher quantity**, then we are justified in saying that x is of a **higher quality** than y.

1. Provide examples to support the **non-standard view**.

   (a) Example 1:

   - Wine X (same quantity) $\succ$ Wine Y (same quantity)
   - Wine X (slightly **lower-quantity**) $\succ$ Wine Y (slightly **higher-quantity**)
   - Appealing to the **higher quality** of Wine X is the most natural way for us to explain these **preference order relations**.

   i. Example 2:

   - Wine X (**lower-quantity**) $\succ$ Wine Y (some **higher quantity** $\leq n$ units)
   - Wine Z (**lower-quantity**) $\succ$ Wine Y (some **higher quantity** $\leq m$ units, where $m > n$)

   ii. Example 3:

   - Wine X (**1 glass**) $\succ$ Wine Y (**ANY quantity**)

Therefore, the **standard view** is a **special case** of the **non-standard view**.

10.10 Haydn and the oyster thought experiment

- Suppose you are a soul waiting to be allocated life on Earth.
- The angel offers you a choice between:
  - ϕ_1: Living the life of **Joseph Haydn**. Haydn composed some wonderful music, influenced the evolution of the symphony, was cheerful and popular, travelled, and enjoyed field sports.
  - ϕ_2: Living the life of an **oyster**. The **oyster's life** consists only of **mild sensual pleasure**. However, the **oyster's life** can be as long as you like.

10.10.1 Bentham's hedonic calculus

The main part of the differences between the options in this thought experiment is the duration in which the **pain or pleasure lasts**.

- Let m be the duration of the **oyster's life**.
- Let n be a **threshold for sufficiency**.
- m is a **sufficiently long duration** if and only if $m \geq n$.
- If the **oyster's life** is **sufficiently long** $m \geq n$, then:

  ϕ_2(where the oyter's life is m years) $\succ$ ϕ_1(where Haydn's life is 77 years)

However, this cannot be right.

10.10.2 Response

- We could ditch **quantitative hedonism** in favour of **qualitative hedonism**.
- **Qualitative hedonism** gives us automatic reasons to favour **higher-quality pleasure** (such as the life of Joseph Haydn) over **lower-quality pleasures** (such as an oyster's life).
- However, according to the **non-standard view of qualitative hedonism**, we may sometimes have to make **trade-offs** between a **lower quantity of a higher-quality pleasure** and a **higher quantity of a lower-quality pleasure**.
- If the **oyster's life** is **sufficiently long** $m \geq n$, then a **trade-off** may have to be made.
- Hence, ϕ_2 (where the oyster's life, though **lower-quality**, is **sufficiently long**) $\succ$ ϕ_1 (where Haydn's life, though **higher-quality**, is **insufficiently long**)

10.11 Utility monster thought experiment

- Suppose that a **utility monster** is a hypothetical entity that is a **highly efficient consumer of resources**.
- The **utility monster** gains vast amounts of **pleasure (happiness)** from very small quantities of a particular resource.
- You have to choose between:
  - ϕ_1: Satisfying the needs of the ordinary human beings.
  - ϕ_2: Satisfying the needs of the **utility monster**.
- Given the existence of this **utility monster** and the **core consequentialist commitment** to **maximising the good**, it seems that we ought to neglect ϕ_1 in favour of ϕ_2.
- However, this cannot be right.

10.12 Against hedonism

- **Consequentialists** will agree that:
  - ϕ_1 (living the life of **Joseph Haydn**) $\succ$ ϕ_2 (living the life of an **oyster**), however long the oyster's life might be.
  - ϕ_1 (satisfying the needs of ordinary human beings) $\succ$ ϕ_2 (satisfying the needs of the **utility monster**), however **efficient** a **consumer of resources** the **utility monster** might be.
- **Consequentialists** will typically not give up the **core consequentialist commitment** to **maximising the good**.
- However, **consequentialists** may give up the **hedonistic** theory of the good.
- The good is defined as that which **facilitates human pleasure and happiness**.
- **Ethical hedonism** is a key element in the **hedonistic** theory of the good, as only **pleasure (happiness)** has **value** and only **pain (unhappiness)** has **disvalue**.

10.12.1 Objections to ethical hedonism

Objection	Elaboration
Deny **value monism**	**Ethical hedonism** implies **value monism**, where n denotes the number of **intrinsic goods**. **Value monism** is when $n = 1$, while **value pluralism** is when $n > 1$. However, we could maintain that while we ought to **maximise the good**, the **good** include far more than can be reduced to **pleasure (happiness)**. Hence, **pleasure (happiness)** is not the only **intrinsic good**. Other **intrinsic goods** include **beauty, friendship**, and so on. This leads to **value pluralism**.
Deny that certain **pleasure states** are **intrinsically good**. Problems with both **quantitative hedonism** and **qualitative hedonism**.	A **sadist whipping her victim** or an **addict on drugs** might experience **pleasure (happiness)**. However, these **pleasure states** are not **intrinsically valuable**. **Quantitative hedonism** might lead us to overvalue **bestial, unsophisticated, lower-quality**, and **debauched pleasures** (the push-pin versus poetry example). At the same time, **qualitative hedonism** faces a **dilemma**. Either **qualitative hedonism** will turn out to be **quantitative hedonism** in disguise, or **qualitative hedonism** will be **inconsistent**.
Alternative **theories of the good** are superior to the **hedonistic** theory of the good.	**Pleasure** and **pain sensations**, inside the **heads of human beings**, are **difficult to measure**. By contrast, **well-being** can be defined in terms of **preference fulfilment** or **desire satisfaction**, giving rise to alternative **theories of the good**.

10.12.2 Alternative theories of the good

- Present desire satisfaction theory: The **good** is defined as that which **facilitates the satisfaction of our current desires**. We also need a **fitting attitude account**, according to which what is **desired** is closely linked with what is **good**.
- Comprehensive desire satisfaction theory: The **good** is defined as that which **facilitates the satisfaction of desires over the course of our life**.
- Informed desire satisfaction theory: The **good** is defined as that which **facilitates the satisfaction of the desires we would have** if we were **fully informed of all the relevant facts**.
- Objective list theory: The **good** is defined as that which does not consist merely in either **pleasurable experience (hedonistic theory of the good)** or **desire satisfaction (desire satisfaction theory)**.

10.12.3 Desire satisfaction theory

- Desire satisfaction theory is also known as **preference fulfilment theory**.
- We could get agents to **rank-order their preferences** and develop **utility functions** on their behalf.
- **Preference utilitarianism** is based on **desire satisfaction theory**.
- Relative to the agents' reported and **rank-ordered preferences** and these **utility functions**, we may be able to derive the following:
  - $U(\phi_1) > U(\phi_2)$, however long the oyster's life might be. Hence, ϕ_1 (living the life of **Joseph Haydn**) $\succ \phi_2$ (living the life of an **oyster**).
  - $U(\phi_1) > U(\phi_2)$, however **efficient** a **consumer of resources** the **utility monster** might be. Hence, ϕ_1 (satisfying the needs of ordinary human beings) $\succ \phi_2$ (satisfying the needs of the **utility monster**).

10.12.4 Minimisation of the violation of rights

- The minimisation of the violation of rights could be identified as another relevant end for **consequentialism**.
- Human beings have the **right to resources** that allow their needs to be satisfied.
- Their **right to resources** function as **side constraints** on the pursuit of **good consequences**.
- This gives rise to a **utilitarianism of rights**.
- Therefore, ϕ_1 (satisfying the needs of ordinary human beings) $\succ$ ϕ_2 (satisfying the needs of the **utility monster**).

10.13 Formal definition of consequentialism

- The term "consequentialism" was first introduced by G. E. M. Anscombe in "Modern moral philosophy" (1958).
- According to consequentialism, only the **consequences of actions** matter, whereas the **intentions behind actions** are unimportant.

The elements of consequentialism include:

- A **core consequentialist commitment**, which is the **maximisation of the good**.
- A **collection of theories of the good**.
  - **Egoistic**: The good is defined as that which **facilitates self-interest**.
  - **Hedonistic**: The good is defined as that which **facilitates human pleasure and happiness**.
  - **Desire satisfaction theory**: The good is defined as that which **facilitates the satisfaction of our desire**.
  - **Objective list theory**: The good is defined as that which does not consist merely in either **pleasurable experiences (hedonistic theory of the good)** or **desire satisfaction (desire satisfaction theory)**.

10.13.1 Semantics of consequentialism

- The semantics of consequentialism can be interpreted in terms of a **semantics of possible worlds**.
- To **maximise the good** is to make the **world**, the **sum of all things**, as **good as it can be**.
- Alternatively, to **maximise the good** is to act to bring about the **best possible world** or all those worlds that can be brought about.
- An **agent** (j, k, and so on) is a **being capable of actions** that are **apt for moral evaluation**.
- Each **action** ϕ_i brings about a **possible world** w_i.
- Hence, an associated set of **possible worlds** $w_1, w_2, \ldots, w_n$ is brought about by the alternative courses of action $\phi_1, \phi_2, \ldots, \phi_n$.
- **Possible worlds** $w_1, w_2, \ldots, w_n$ are alternatives between which an **agent** j must choose.

10.13.2 Definitions

- Let Φ denote the set of alternative courses of action $\phi_1, \phi_2, \ldots, \phi_n$.

$$\Phi = \{\phi_1, \phi_2, \ldots, \phi_n\}$$

- Let W denote the set of **possible worlds** $w_1, w_2, \ldots, w_n$, brought about by the actions in Φ.

$$W = \{w_1, w_2, \ldots, w_n\}$$

- Let A denote the association of members of Φ with their associated **possible worlds** in W.

$$A = \{\phi_1 - w_1, \phi_2 - w_2, \ldots, \phi_n - w_2\}$$

- Let $< j, A >$ denote a **choice situation**, where j is the agent, A is the set of associations.
- Let R denote a **rightness function** that assigns to each **choice situation** $< j, A >$, a **subset of** A.

$$< j, A > \xrightarrow{R} R_j(A)$$

$$\{\phi_1 - w_1, \phi_2 - w_2, \ldots, \phi_n - w_n\} \xrightarrow{R} \{\phi_1 - w_1(\text{good}), \cancel{\phi_2 - w_2(\neg \text{good})}, \ldots, \cancel{\phi_n - w_n(\neg \text{good})}\}$$

10.13.3 Conditions

Let T denote a **theory of good**, representable in terms of a **complete order** $\leq$ on the **set** W **of all possible worlds**.

Condition	Description	Formal representation
Reflexivity	x is **at least as good as itself**, according to T.	$x \leq x$
Transitivity	If y is **at least as good as** x and z is **at least as good as** y, then z is **at least as good as** x, according to T.	$((x \leq y) \wedge (y \leq z)) \rightarrow (x \leq z)$
Completeness	It is either the case that y is **at least as good as** x or that x is **at least as good as** y, according to T.	$(x \leq y) \vee (y \leq x)$

10.13.4 Axioms of consequentialism

1. AN or agent neutrality. Consequentialism is **agent-neutral**. For any two agents j and k:

   $$R_j(A) = R_k(A)$$

   Essentially, it doesn't matter who or what the agent is, the right thing to do will always be the same.

2. NMD or **no moral dilemmas**. There are **no moral dilemmas** under **consequentialism**.

   $$R_j(A) \neq \emptyset$$

3. DM or **dominance**. x **dominates** y under **consequentialism**.

   - For any two possible worlds x and y, an agent j, and any two possible choice situations $< j, A >$ and $< j, B >$:
   - Suppose that $\{x, y\} \subseteq A \cap B, x \in R_j(A)$ and $y \notin R_j(A)$.

     $$\therefore y \in R_j(B)$$

   - Essentially, what this axiom states is that once you have determined that one action is better than another action in a decision, the other action that you have rejected cannot be the best action in the next decision you make, because there is already a better action available.

10.13.5 Elements of consequentialism

1. The **agent** j.
2. A set Φ of alternative courses of action $\phi_1, \phi_2, \ldots, \phi_n$.
3. A set W of **possible worlds** $w_1, w_2, \ldots, w_n$ brought about by the actions in Φ.
4. An association A between members of Φ and their associated **possible worlds** in W.
5. The **choice situation** $< j, A >$.
6. The **rightness function** R.
7. The **theory of the good** T, defined in terms of the following conditions:

$$\text{Reflexivity: } x \leq x$$

$$\text{Transitivity: } ((x \leq y) \wedge (y \leq z)) \rightarrow (x \leq z)$$

$$\text{Completeness: } (x \leq y) \vee (y \leq x)$$

8. The **axioms of consequentialism**:

$$(AN)\ R_j(A) = R_k(A)$$

$$(NMD)\ R_j(A) \neq \emptyset$$

(DM) Suppose that $\{x, y\} \subseteq A \cap B, x \in R_j(A)$ and $y \notin R_j(A)$:

$$\therefore y \notin R_j(B)$$

10.14 Violation of consequentialist axioms

It is possible for at least some of the **axioms of consequentialism** to be violated.

10.14.1 Violating the agent neutrality axiom

- **Agent-relative consequentialism** combines the **core consequentialist commitment** (each agent ought to **maximise the good**) with an **agent-relative axiology**.
- The correct evaluation of the **complete order** $\leq$ on the **set** W **of all possible worlds** may vary from agent to agent.
- If there is an **agent-relative axiology**, then $\Diamond(R_j(A) \neq R_k(A))$.
- Hence, AN **(agent neutrality)** may be violated.

10.14.2 Violating the no moral dilemmas axiom

- If the **theory of the good** T is defined in terms of **completeness**, then for any two possible worlds x and y, either $x > y, x = y$ or $y > x$.
- The **good** must be represented by a **complete order** $\leq$.
- This rules out **incommensurability**.
- However, if $\leq$ is **incomplete**, then we could allow for **incommensurability between at least some possible worlds**.
- Therefore, this would allow **moral dilemmas** within **consequentialism**.

$$R_j(A) = \emptyset$$

- Therefore, NMD (**no moral dilemmas**) will be violated.
- The possibility of **moral dilemmas** within **consequentialism** might require us to drop NMD.

10.14.3 Violating the dominance axiom

- **Satisficing consequentialism** relaxes the **maximising element** in **consequentialism**.
- According to **satisficing theories**, it is sometimes permissible to do something than is **worse than the best**, provided that it is **good enough**.
- Hence, DM (**dominance**) will be violated.
- Relative to the choice situations $< j, B >$, the **satisficing threshold** is **sufficiently low**.
- Therefore, $x \in R_j(B)$ and $y \in R_j(B)$ (both x and y are **good enough**).
- Hence, x does not **dominate** y.
- DM (**dominance**) will be violated.

10.15 Driver's strategy

10.15.1 Move 1

Identify objections to **consequentialism** raised by the **standard feminist view**.

- Objection 1: **Consequentialism** is **too demanding** of the individual.
- Objection 2: **Consequentialism** is **neglectful of an agent's special obligations** to family and friends.

10.15.2 Move 2

Identify the problems of **consequentialism**.

- Problem 1: **Impartiality** - we should **maximise the good, impartially considered**.
- Problem 2: **Demandingness** - **consequentialism** appears to make **theoretically unlimited demands**.

10.15.3 Move 3

Introduce **conceptual distinctions**.

- The angel of the house **is partial to her domestic sphere** and **self-sacrificing**.
- The angel of the world **is impartial** and **self-sacrificing**.
- **Moral self-sufficiency**: There is a concern with promoting an **internalist standard of moral worth**, as our wills are sufficient to ground our moral worth, and we do not need to look externally for the source of this value.
- **Ethical self-sufficiency**: There is a concern with promoting a **risk aversion standard**.
  - The best life needs to incorporate a **respect for personal space**.
  - There is space in **ethics** for **self-perfection, personal projects, goals,** and **aspirations**.

10.15.4 Move 4

Articulate a **sophisticated consequentialism**.

- The aim here is to avoid the **angel of the world** and **ethical self-sufficiency** (Schopenhauer's ideal).
- Introduce further **conceptual distinctions** between:
    1. A **decision procedure** and a **criterion for rightness of actions**. A **decision procedure** may interfere with the actual production of the good, as determined by some **criterion of rightness**.
    2. **Truth conditions** and **acceptance conditions**. We could hold that **consequentialism** is true, although we do not accept it in certain areas of our life.

10.15.5 Move 5

Anticipate objections to **sophisticated consequentialism**.

- Objection: **Moral schizophrenia**.
- Response: **Moral schizophrenia** does not speak against the **truth** of a theory.
- It merely shows that the theory is **difficult to apply** as a **decision procedure**.
- A **theory** may be **true** yet difficult to apply.
- The **impartial good** will be promoted if the **angel of the world** cultivates **special relationships** and becomes less of an all-encompassing angel.
- Therefore, we now have a **norm of partiality**, which means we ought to show preference for our family and friends.

10.15.6 Conclusion

As a **sophisticated consequentialist**, one can be **both a consequentialist and a feminist**. One can be responsive to both **feminist concerns about partiality** and the **demands of morality**.

11 Deontology

11.1 Hiroshima and Nagasaki

- Should the atomic bombings of Hiroshima and Nagasaki have been carried out or not?
- Let ϕ denote the action of carrying out the atomic bombings of Hiroshima and Nagasaki.
- Since ϕ-ing has brought about the **best possible consequences** (the end of WWII and world peace), ϕ-ing should have been carried out.

11.1.1 Argument in favour of ϕ-ing

- Only the **consequences of actions** matter, whereas the **intentions behind actions** are unimportant.
- There are no **side constraints** under **consequentialism**.
- ϕ-ing would be **morally right** due to the **maximisation of the good**.
- Therefore, the **evaluative claim** under **consequentialism** is that we ought to ϕ.

11.1.2 Argument not in favour of ϕ-ing

- **Side constraints** exist under **deontology** ϕ-ing would be **morally wrong** due to the **violation** of certain **side constraints**.
- Therefore, the **deontic verdict** under **deontology** is that we ought not to ϕ.

11.2 Textbook view

According to the **textbook view**:

- **Consequentialism** is an **agent-neutral theory**, whereas **deontology** is an **agent-relative theory**.
- The **textbook view** has also been describe as the **standard method** for drawing the distinction between **consequentialism** and **deontology**.

The **textbook view** implies the following:

Deontology	Consequentialism
Deontology may give **different agents** j, k, etc. **different aims**.	**Consequentialism** gives **different agents** j, k, etc. the **same aim**: select the action ϕ_i that **maximises the good**.
Deontology introduces **agent-relative side constraints**.	**Consequentialism** does not observe any **agent-relative side constraints**.
$\Diamond(R_j(A) \neq R_k(A))$ with an **agent-relative axiology**. Hence, AN (**agent neutrality**) may be violated.	(AN or **agent neutrality**) $R_j(A) = R_k(A)$ AN is an **axiom** under **consequentialism**.
We can give **higher weight** to our interests, projects, and concerns.	We must give **equal weight** to our interests and the interests of others.
We might have certain **special obligations** (for instance, **obligations** that parents have to their own children), not shared by other agents.	We ought to be **impartial**.
We end up with **agent-relative reasons for action**.	We end with **agent-neutral reasons for action**.

11.2.1 Exceptions to the textbook view

1. **Ethical egoism**
    - **Ethical egoism** is supported by an **egoistic** theory of the good.
    - The **good** is defined that which **facilitates self-interest**.
    - Since what counts as **self-interest** may differ from agent to agent, **ethical egoism** may give **different agents** j, k, etc. **different aims**.
    - Therefore, **ethical egoism** is an **agent-relative** version of **consequentialism**.
2. **Agent-neutral deontology**
    - **Agent-neutral deontology** is an **agent-neutral** version of **deontology**.
    - **Agent-neutral deontology** gives **different agents** j, k, etc. the **same aim**.

11.3 Mafia scenario

- Suppose that the mafia are credibly threatening to kill two strangers unless you **kill a third stranger**.
- Let ϕ denote the action of **killing the third stranger**.
- Should we ought to ϕ or $\neg\phi$?

Deontology	Consequentialism
The **deontic verdict** is that we ought not to $\neg\phi$.	The **evaluative claim** is that we ought to ϕ.
Justification for **deontic verdict**. **Side constraints** exists that **prohibit killing** as a **morally wrong** type of action.	**Justification** for **evaluative claim**. We ought to **minimise the number of bad occurrences**.

11.4 Side constraints

There are two types of **side constraints**.

Type of side constraint	Description	Implication
An **agent-relative side constraint**.	Each agent should ensure that **she does not kill**, even if to **prevent more killings by others**.	**Agent-relative side constraints** might give an agent a **special concern** with her own killings.
An **agent-neutral side constraint**.	Each agent should ensure that **no one kills**, even if to **prevent more killings by others**. **Agent-neutral side constraints** may require everyone to **share a moral vision**.	**Agent-neutral side constraints** gives **different agents** j, k, etc. the **same aim**.

11.5 The right and the good

- The **textbook view** uses the **agent-neutral and agent-relative** distinction to distinguish between **conseuqentialism** and **deontology**.
- Another popular approach to distinguishing between **consequentialism** and **deontology** involves an appeal to the **right or good** distinction:

Deontology	Consequentialism
The **right** is **prior** to the **good**.	The **good** is **prior** to the **right**.
Deontology delivers **deontic verdicts** in terms of **what is right**. If an action ϕ is **morally impermissible**, then it is **not right**, no matter how much **good** might be produced by ϕ-ing.	**Consequentialism** delivers **evaluative claims** in terms of **what is good**.

11.6 Action and intention

We can distinguish between different types of **deontology** in terms of the **action or intention** distinction and the **agent versus patient** distinction:

Types of deontology	Description
Action-based deontology	The **right action** is **prior** to the **good consequences**.
Intention-based deontology	The **right intention** is **prior** to the **good consequences**.
Action and intention-based deontology	The **right action** and **right intention** are **prior** to the **good consequences**.
Agent-centred deontology	The primary concern is with the **duties of the agent**.
Patient-centred deontology	The primary concern is with the **rights of the patient**. The **patient** has a **right against being used** as a mere **means for producing good consequences** without her consent.

11.7 Doctrine of double effect

- **Intention-based deontology** is supported by the **doctrine of double effect**.
- The doctrine of double effect states that it is **morally impermissible** for us to **intend evil** (for instance, the **killing or torturing of innocents**).

Possible objections to the **doctrine of double effect**:

1. The distinctions between **intending, foreseeing, risking, predicting** and **causing and allowing** are **conceptually incoherent**.
2. The distinctions invite **manipulation** and can be **exploited**.

11.8 Avoision

- Suppose that there is a **moral prohibition** against ϕ_1-ing.
- However, ϕ_2-ing allows an agent to **bypass, circumvent,** or **duck this moral prohibition**.
- Does ϕ_2-ing count as a **morally permissible avoidance** of the **moral prohibition**?
- Does ϕ_2-ing count as a **morally impermissible evasion** of the **moral prohibitions**?
- This type of **manipulation** in the legal domain has been termed **avoision**.
- The distinctions between **intending, foreseeing, risking, predicting** and **causing and allowing** could give rise to a **moral** version of **avoision**.

11.9 Trolley dilemma

- According to the **trolley dilemma**, a **runaway tram or trolley** is on course to run over and **kill five people on the main track**.
- However, you can intervene, **pull the lever**, and divert the runaway tram or trolley to a **side track, killing just one person**.
- You have to choose between:
  - ϕ: **Pulling the lever** and **killing the one person on the side track** to **save the five people on the main track**.
  - $\neg\phi$: **Refraining from pulling the lever** and **letting the five people on the main track die**.
- Should you **pull the lever** and **kill one** to **save five** ϕ or **refrain from doing so** $\neg\phi$?

11.9.1 Positions

- **Consequentialism** recommends **pulling the lever** and **killing one to save five** (ϕ).
  - ϕ-ing will **minimise the number of bad occurrences**.
  - We ought to **minimise the number of bad occurrences**.
  - Hence, we ought to ϕ.
- **Action-based deontology** recommends **refraining from pulling the lever** ($\neg\phi$).
  - ϕ-ing is a **morally impermissible action**.
  - $\neg\phi$-ing is a **morally permissible action**.
  - Hence, we ought to $\neg\phi$.
- **Intention-based deontology** recommends two different actions.
  1. **Pull the lever** and **kill one to save five** (ϕ).
     - ϕ-ing may be accompanied by the **intention of saving five**.
     - We merely **risk, foresee, or predict** that ϕ-ing will have the consequence of one innocent person being killed.
     - Hence, we ought to ϕ.
  2. **Refrain from pulling the lever** ($\neg\phi$).
     - ϕ-ing may be accompanied by the **intention of killing one innocent person**.
     - We should not **intend evil**.
     - Hence, we ought to $\neg\phi$.
- **Patient-centred deontology** recommends **refraining from pulling the lever** ($\neg\phi$)
  - **That one person on the side track** has a **right against being used** as a mere **means for producing good consequences** without their consent.
  - Hence, we ought to $\neg\phi$.

11.10 Siamese twins

- Suppose that **Siamese twins** are conjoined such that **both will die** unless the **organs of one are given to the other** via an operation that kills the first.
- You have to choose between:
  - (ϕ): Performing the **life-saving operation**.
  - ($\neg\phi$): **Refraining from** performing the **life-saving operation**.

11.10.1 Positions

- **Intention-based deontology** recommends **performing the life-saving operation** (ϕ)
  - We **intend** to save the first twin in a **life-saving operation**.
  - We do not **intend** but merely **foresee** the death of the second twin in that operation.
  - Hence, we ought to ϕ, justified by the use of **foreseeing and intending** distinction.
- **Action-based deontology** recommends **performing the life-saving operation** (ϕ)
  - We **cause the first twin to be saved** in a **life-saving operation**.
  - We do not **cause** but merely **allow** the death of the second twin in that operation.
  - Hence, we ought to ϕ, justified by the use of the **causing and allowing** distinction.
- **Patient-centred deontology** recommends **refraining from performing the operation** ($\neg\phi$)
  - Each twin has a **right against being used** as a mere **means for producing good consequences** without her consent.
  - Hence, we ought to $\neg\phi$.

11.11 Divine command theory

- **Divine command theory** is a form of **deontology**.
- In **divine command theory**, the **side constraints** are provided by the **divine commands of God**.
- Examples of **divine commands** and religious precepts supporting the **side constraint against murder**:
    1. Thou shalt not kill, from Exodus 20:13. This is one of the **10 Commandments** in **Judaism** and **Christianity**.
    2. I undertake the precept to **refrain from destroying living creatures**. This is one of the **five precepts in Buddhism**.
    3. One ought to **avoid harming (or desiring to harm) any living being** in thought, word or deed. This is the **doctrine of Ahimsa** in **Jainism, Hinduism,** and **Buddhism**.
- These religious commandments, precepts, and doctrines **prohibit killing** as a **morally wrong** type of action and function as **side constraints**.

11.11.1 Definition

- **Divine command theory** is the view according to which:
    - **Morality** is somehow dependent upon **God**.
    - **Moral obligation** consists in **obedience to God's commands**.
- **Divine command theory** consists of at least some of the following claims:
    1. **Morality** is the ultimately based on the **commands of God**.
    2. **Morality** is ultimately based on the **character of God**.
    3. **Moral obligations** are identical with **divine commands**.
    4. **Moral obligations** are created by **divine commands**.
- **Divine command theory** has to address the **Euthyphro dilemma**.

11.11.2 Euthyphro dilemma

- Socrates' original question to Euthyphro: Is the **pious** loved by the gods because it is pious, or is it pious because it is loved by the gods?
- Revised version of Socrates' question: Is an action **divinely commanded** by God because it is **morally right**, or is an action **morally right** because it is **divinely commanded** by God?
- Horn 1: An **action** is **divinely commanded** by God because it is **morally right**. Problems associated with this horn:
    1. God would no longer be the **author of morality**. Rather, God would be merely a **being capable of recognising right and wrong**.
    2. God would not be the **sovereign of morality**. Rather, God would be a mere **subject of morality**.
- Horn 2: An **action** is **morally right** because it is **divinely commanded** by God. Problems associated with this horn:
    1. Nothing guarantees that what God would **divinely command** would be in accordance with what **morality** would prescribe. If God commands us to do something **morally reprehensible**, then the **morally reprehensible** action would be **morally right**.
    2. The **foundations of morality** become **arbitrary**. **Morally reprehensible actions** can become **morally obligatory** if **divinely commanded** by God.

11.12 Modified divine command theory

- According to **standard divine command theory**: "ϕ-ing is **morally wrong**" means "ϕ-ing is contrary to the **divine commands** of God."
- According to the **modified divine command theory**: "ϕ-ing is **morally wrong**" means "ϕ-ing is contrary to the **divine commands** of a **loving** God."
- **Modified divine command theory** holds that God has a particular character.
- God has the character of **loving his human creatures**.
- Therefore, while it is **logically possible** for God to command **morally reprehensible** action, it is not actually something that God would do, given his character.

11.12.1 How it fixes the issues with Euthyphro's dilemma

Horn 1:

- God is the **subject** rather than the **sovereign of morality**.
- Under **modified divine command theory**, God remains the **source of morality**.
- **Morality** is grounded in the **loving** character of God.

Horn 2:

- The **foundations of morality** become **arbitrary**.
- Under **modified divine command theory**, not any **divine command** goes.
- **Divine commands** remain rooted in the unchanging **loving** and **omnibenevolent** character of God.

Standard divine command theory is pierced by either **horn 1** and **horn 2**. However, the **modified divine command theory** appears to fare better against the **Euthyphro dilemma**.

11.13 Plato

11.13.1 Context

- **Socrates** is being accused of **corrupting the youth of Athens**.
- **Socrates** is about to be put to trial and thereafter found **guilty** of both:
  1. **Corrupting the minds of the youth of Athens**.
  2. **Impiety** ("not believing in the Gods of the state").
- **Euthyphro** is **prosecuting his father** for **murder**.
- Both Socrates and Euthyphro try to arrive at a definition of **piety** or **holiness**.
- Euthyphro's modified definition of **piety** or **holiness** (suggested by Socrates):
  - Holy is defined as what all the gods love.
  - Unholy is defined as what all the gods hate.
  - $\neg$ (Holy $\vee$ Unholy) is defined as what some gods love and some other gods hate.
  - (Holy $\vee$ Unholy) is defined as what some gods love and some other gods hate.

11.13.2 The Euthyphro's dilemma

- The **Euthyphro dilemma**: Is the **holy** loved by the gods because it is holy? Or is it holy because it is loved by the gods?
- Revised version of the **Euthyphro dilemma**: Is what is **morally good** commanded by God or the gods because it is **morally good**? Or is it **morally good** because it is commanded by God or the gods?
- **Species-genus** distinction:
  - The **species** is a **part** of the **genus (whole)**.
  - **Reverence (species)** is a **part** of **fear (genus)**.
  - **Piety** or **holiness (species)** is a **part** of **justice (genus)**.
- According to **Euthyphro**: The part of **justice (genus)** that is **holy (species)** involves **ministering to** (looking after or **taking care** of) the gods.

11.13.3 Socrates' argument by analogy

Source	Target
We **take care of animals**.	The part of **justice (genus)** that is **holy (species)** involves **taking care of the gods** (Euthyphro's claims).
Taking care of horses, dogs, and **cattle** in horsetraining, huntsmanship, and herdsmanship **benefits and improves horses, dogs**, and **cattle**. You make these animals **better**.	You do not make the gods **better** when you do something **pious** or **holy**.

Therefore, Euthyphro's claims are problematic.

11.13.4 Conclusion

- The **pious** or **holy** is in fact **something other than what is acceptable to the gods**.
- Something's being acceptable to the gods is merely an **attribute** of **piety** and not part of its **defining characteristics**.
- Therefore, Socrates still does not have a **definition of piety** or **holiness**.

11.14 Secular deontology

Deontology may be **religious** or **secular** in nature.

11.14.1 Religious form of deontology

- All **duties and obligations** are generated by **religious precepts, commandments**, and **doctrines**.
- An example is **divine command theory**, which has **religious precepts, commandments**, and **doctrines** that are of **divine origin**.

11.14.2 Secular form of deontology

- All **duties and obligations** are generated by **secular principles**.
- An example is **deontological monism**, where all **duties and obligations** are generated by a **single secular principle**.
- Another example is **deontological pluralism**, where all **duties and obligations** are generated by **multiple secular principles**.

11.15 Kantian deontology

- **Kantian deontology** is an example of a **secular** form of **deontology**.
- All **duties and obligations** are generated by a **single secular principle**.
- This **single secular principle** is known as the **categorical imperative**.

11.15.1 Types of imperatives

1. Hypothetical imperative
    - A **command** that applies to us **conditionally**.
    - **Hypothetical imperatives** contain conditional clauses that can be explicit or elided.
    - **Hypothetical imperatives** require us to exercise our wills in a certain way, given that we have **antecedently** willed a particular end.
    - An example is, "If you are happy, and you know it, clap your hands".
2. Categorical imperative
    - A **command** that applies to us **unconditionally**, regardless of what we might want.
    - **Categorical imperatives** do not have any **conditional clauses**, that can be **explicit** or **elided**.
    - An example is, "Thou shalt not kill."

11.15.2 Categorical imperative formulations

1. The **universal law of nature** formula.
2. The **humanity** formula.
3. The **autonomy** formula.
4. The **kingdom of ends** formula.

Formulations 1 to 4 are supposed to be **equivalent**, hence **Kantian deontology** is also an example of **deontology monism**.

11.15.3 Universal law of nature formula

- The **universal law of nature** states:
  Act only in accordance with that **maxim** that you can at the same time will that it become **universal law**.

- Decision procedure associated with the **universal law of nature** formula:

  1. Consider a particular action ϕ (for instance, stealing).
  2. Determine a **maxim** governing ϕ (for instance, I will steal for pleasure).
  3. **Universalise** that **maxim** and consider the implications (for instance, everyone ought to steal for pleasure).
  4. Consider whether the **universalised** version of the **maxim** is **contradiction-free**. Could the **universalised** version of the **maxim** function as a **law of nature**?

- Example: Suppose I borrow money from you and promise to return the amount eventually, but I know that I will not return the amount.

11.15.4 Detailed steps for the universal law of nature formula

1. Consider a particular action ϕ.
    - Here, ϕ-ing denotes the action of **borrowing money and promising to repay**, despite knowing that **this will never be done**.
2. Determine a **maxim** governing ϕ.
    - Maxim: Whenever I am low on cash, I will **borrow money and promise to repay**, despite knowing that **this will never be done**.
3. **Universalise** that **maxim** and consider the implications.
    - **Universalised** version of maxim: Whenever anyone is low on cash, she ought to **borrow money and promise to repay**, despite knowing that **this will never be done**.
    - However, if this deceit is **universalised**, there will be no institution of **promising** to begin with.
4. Consider whether the **universalised** version of the **maxim** is **contradiction-free**.
    - A **contradiction** arises, as we cannot **make promises** if there is no institution of **promising** to begin with.
    - Therefore, I will be unable to make my **promise** to begin with.

11.15.5 Humanity formula

- The **humanity** formula states:
  Act in such a way that you always treat **humanity**, whether in your own person or in the person of any other, never simply as a **means**, but always at the same time as an **end**.
- Implications:
    - We ought to **respect the humanity in persons**.
    - We ought to refrain from treating persons as **mere instruments**.

11.15.6 Autonomy formula

- The **autonomy** formula: Act in such a way that your **will** can regard itself at the same time as **making universal law through its maxims**.
- Implications:
  - We ought to act as **universal lawgivers** or **legislators**.
  - We ought to consider whether our intended maxims are worthy of our status as **shapers of the world**.
  - As **rational agents**, we are the very **source of the authority** for the **moral laws** that bind us.

11.15.7 Kingdom of ends formula

- The **kingdom of ends** formula: Act in accordance with the **maxims** of a member giving **universal laws** for a **merely possible kingdom of ends**.
- Implications:
  - We ought to consider whether our intended maxims will earn **acceptance** by a **community of fully rational agents** in a **kingdom of ends**.
  - Just as **human beings** are **ends in themselves**, we are also a **kingdom of ends** or a **moral community**.

11.16 Organ transplant thought experiment

- Suppose that **five mortally ill patients** are at a hospital.
- They will soon die without an organ transplant.
- Each of these patients requires a different organ to be transplanted (for instance, a **heart** for the first patient, a **kidney** for the second patient, a **liver** for the third patient, and so on).
- At the same time, a **sixth healthy patient** is undergoing a routine check-up at the same hospital.
- The only medical means of **saving the five mortally ill patients** would be to **kill the sixth healthy patient** and transplant his healthy organs into the five other patients.
- You have to choose between:
  - (ϕ): Killing the **one healthy patient** to **save the five mortally ill patients**.
  - ($\neg\phi$): Refraining from killing the **one healthy patient**, even at the cost of the **five mortally ill patients** dying.
- Should you kill the **one healthy patient** to **save the five mortally ill patients** (ϕ) or refrain from doing so ($\neg\phi$)?

11.16.1 Consequentialist position

- Consequentialism recommends **killing the one healthy patient** to **save the five mortally ill patients** (ϕ).
- Justification:
  - ϕ-ing will **maximise the good**.
  - Certain **theories of the good** might require us to ϕ.
  - Therefore, we ought to ϕ.

11.16.2 Deontological position

- The deontological position recommends refraining from killing the one healthy patient $\neg\phi$.
- Justification:
  - ϕ-ing is a **morally impermissible action**.
  - $\neg\phi$-ing is a **morally permissible action**.
  - The **humanity** formula introduces **side constraints** against ϕ-ing.
  - ϕ-ing requires that you treat the **sixth healthy patient** as a mere **means** or **conduit** for **five useful and life-saving organs**.
  - Hence, we ought to $\neg\phi$.

11.16.3 Permissivity

For consequentialism:

- The **evaluative claim** of **consequentialism**:
  - ϕ-ing is not merely **morally permissible**.
  - ϕ-ing is also **morally obligatory**, we ought to **maximise the good (core consequentialist commitment)**
  - Hence, **consequentialism** appears to be **overpermissive**.

For deontology:

- The **deontic verdict** of **deontology**:
  - ϕ-ing is **morally impermissible**.
  - ϕ-ing violates certain **side constraints** (including the **humanity** formula)
  - Hence, **deontology** is not **overpermissive**.

Note that the **organ transplant** thought experiment and the **trolley dilemma**:

- (Trolley dilemma): **Pull the lever** and **kill one** to **save five** (ϕ) or **refrain from doing so** ($\neg\phi$)?
- (Organ transplant): **Kill the sixth healthy patient** to **save five** ϕ or **refrain from doing so** ($\neg\phi$)?

11.17 Rossian deontology

- **Kantian deontology** is an example of **deontological monism**.
- All **duties and obligations** are generated by a **single secular principle**.
- By contrast, **Rossian deontology** is an example of **deontological pluralism**.
- All **duties and obligations** are generated by **multiple secular principle**.
- Furthermore, these **multiple secular principles** cannot be reduced to a **single master or mistress principle**.

11.17.1 Principles

- Fidelity: We have a duty to **keep our promises**.
- Reparation: We have a duty to **right our previous wrongs**.
- Gratitude: We have a duty to **return services** to those from whom we have accepted benefits in the past.
- Beneficence: We have a duty to promote a **maximum of aggregate good**.
- Nonmaleficence: We have a duty to **refrain from harming others**.

11.17.2 Prima facie duties

- **Rossian principles** specify **aspects of a situation** that **count morally** in favour of an action ϕ.
- For instance, an action ϕ that allows us to **right our previous wrongs (reparation) counts morally** in favour of ϕ-ing.
- **Rossian principles** yield **prima facie duties**.
- However, **prima facie duties** can be **outweighed** by other **prima facie duties**.
- Furthermore, **moral dilemmas** may arise when **prima facie duties conflict** with each other.
- What the agent ought to choose under **Rossian deontology** is that action ϕ_i: Of all those possible for the agent in the circumstances, that has the **greatest balance of prima facie rightness**, in those respects in which they are **prima facie right**, over their **prima facie wrongness**, in those respects in which they are **prima facie wrong**.

11.17.3 Conflicts between prima facie duties

- **Rossian deontology** maintains that any **conflict between prima facie duties** in a particular situation can be resolved by relying on **intuition**.
- We can rely on our **crystal-clear intuitions** in both **mathematics** and **ethics** to build up all we can know about the **nature of numbers** and the **nature of duty**.
- Our **mathematical knowledge** and our **moral knowledge** are **self-evident**.
- When we assume that AB and CD are parallel, i.e. $AB \parallel CD$, then AB and CD do not meet except at infinity in a Euclidean plane is self-evident by **mathematical intuition**.
- $\neg Pp$, where p denotes the action of **genocide** is **self-evident** by **moral intuition**.
- The **actual duty** is what the agent is left with after she has weighed up all the **conflicting prima facie duties** in the manner prescribed by Ross.

11.17.4 Issues with intuition to resolve conflicts

1. How do we even identify the **prima facie duties** that are involved in a particular situation?
2. What are the **criteria** according to which we **rank and compare prima facie duties**, in order to arrive at the **greatest balance of prima facie rightness over prima facie wrongness** (as prescribed by Ross) that will guide us to our **actual duty**?
3. What if our **intuitions** are **wrong or misguided**?

11.18 W.D.

- W.D. is a machine-based implementation of **Rossian deontology**.
- Jeremy and W.D. are important developments in the domain of **machine ethics**.
- **Machine ethics** is broadly concerned with ensuring that the **behaviour of machines** is **ethically acceptable**.
- The **relations** to be learnt by W.D. are represented as **first-order horn clauses** of the form:

  $$H \leftarrow (L_1 \wedge L_2 \wedge \cdots \wedge L_n)$$

  Where:

  - H is a positive literal H
  - $\leftarrow$ means implication
  - $(L_1 \wedge L_2 \wedge \cdots \wedge L_n)$ is a universally quantified conjunctions of positive literals L_i

- W.D. uses **inductive logic programming** to learn the **supersedes relation**:

  $$\text{supersedes}(\phi_1, \phi_2)$$

  This means that action ϕ_1 is preferred over action ϕ_2 in a particular situation.

11.18.1 Favours relation

The **favours relation** is a **4-ary relation** that is used as a **specifying operation** to aid the supersedes relation:

$$\text{favours}(1 \text{ or } 2, D_{\phi_1}, D_{\phi_2}, R)$$

Where:

- 1 or 2 signifies in which action's favour (ϕ_1 or ϕ_2), a given **prima facie duty** lies. The possible values are $\{1, 2\}$.
- D_{ϕ_1} signifies ϕ_1's **intensity value** for a particular **prima facie duty**. The possible values are $\{-2, -1, 0, +1, +2\}$.
- D_{ϕ_2} signifies ϕ_2's **intensity value** for a particular **prima facie duty**. The possible values are $\{-2, -1, 0, +1, +2\}$.
- R specifies **how far apart** the **intensity values** of these **prima facie duties** can be. The possible values are $\{1, 2, 3, 4\}$.

For any two actions ϕ_1 and ϕ_2:

$$\text{favours}(1, D_{\phi_1}, D_{\phi_2}, R) \leftarrow D_{\phi_1} - D_{\phi_2} \geq R$$

or

$$\text{favours}(2, D_{\phi_2}, D_{\phi_1}, R) \leftarrow (D_{\phi_2} - D_{\phi_2} \geq 0) \wedge (D_{\phi_2} - D_{\phi_2} \leq R)$$

- W.D. begins by making a **hypothesis** about how the **favours relation** supports the **supersedes relation**.
  - Completeness: A **hypothesis** is **complete** if and only if it covers **all the positive cases**.
  - Consistency: A **hypothesis** is **consistent** if and only if it covers **no negative cases**.
- Therefore, a hypothesis could be:
  - (Complete $\wedge$ consistent): **All positive** and **no negative** cases covered.
  - (Complete $\wedge$ $\neg$ consistent): **All positive** and **at least some negative cases** covered.
  - ($\neg$ Complete $\wedge$ consistent): **Not all positive** and **no negative cases** covered.
  - ($\neg$ Complete $\wedge$ $\neg$ consistent): **Not all positive** and **at least some negative cases** covered.

11.18.2 Input information

1. The name of **action** ϕ_1.
2. A rough estimate of the **intensity** of each of the **prima facie duties** satisfied or violated by this action ϕ_i.
    - -2 means a **serious violation** of duty
    - -1 means a **less serious violation** of duty
    - 0 means a duty is **neither satisfied nor violated**
    - +1 means a **minimal satisfaction** of duty
    - +2 means a **maximal satisfaction** of duty

11.18.3 Machine learning procedure

1. When the data entry is complete, W.D. consults its **current version of the supersedes relation**.
   - W.D. determines whether there is any action ϕ_i that **supersedes** all other actions.
   - Formal representation:

$$\begin{aligned}
&\text{supersedes}(\phi_1, \phi_2) \veebar \text{supersedes}(\phi_2, \phi_1) \\
&\text{supersedes}(\phi_1, \phi_2) \veebar \text{supersedes}(\phi_2, \phi_1) \\
&\text{supersedes}(\phi_1, \phi_2) \veebar \text{supersedes}(\phi_2, \phi_1) \\
&\vdots
\end{aligned}$$

2. If this action ϕ_i is discovered, then it will be **output** as the **correct action** to perform. Formal representation:

$$\text{supersedes}(\phi_1, \phi_i) \leftarrow \begin{cases} \text{supersedes}(\phi_1, \phi_2) \\ \text{supersedes}(\phi_1, \phi_3) \\ \text{supersedes}(\phi_1, \phi_4) \\ \text{supersedes}(\phi_1, \phi_5) \\ \quad\vdots \\ \text{supersedes}(\phi_1, \phi_n) \end{cases}$$

3. If no such action exists, then W.D. will seek the **intuitively correct action** from the user.

4. The new information from the user is combined with the **input case** to form a **new training example**. This **training example** is used to refine the **current hypothesis**.

5. The aim of **training W.D.** is to allow it to learn a **new hypothesis** that is a **complete and consistent**, relative to all **input cases**.

11.18.4 Example 1

- We could either **refrain from killing an innocent person** and **using his heart to save another person's life** (ϕ_1) or **kill that person to use his heart to save the other person's life** (ϕ_2).
- The **intuitively correct** response is ϕ_1:
  - For ϕ_1:
    - $\text{Beneficence}_{\phi_1} = -2$
    - $\text{Nonmaleficence}_{\phi_1} = +2$
  - For ϕ_2:
    - $\text{Beneficence}_{\phi_2} = +2$
    - $\text{Nonmaleficence}_{\phi_2} = -2$
- W.D. starts with the **most general hypothesis**: $\text{supersedes}(\phi_1, \phi_2)$.
- The list of **least specific specialisations** for the **favours relation** includes:
  - $\text{favours}(1, \text{fidelity}_{\phi_1}, \text{fidelity}_{\phi_2}, 1)$
  - $\text{favours}(1, \text{reparation}_{\phi_1}, \text{reparation}_{\phi_2}, 1)$
  - $\text{favours}(1, \text{gratitude}_{\phi_1}, \text{gratitude}_{\phi_2}, 1)$
  - $\text{favours}(1, \text{beneficence}_{\phi_1}, \text{beneficence}_{\phi_2}, 1)$
  - $\text{favours}(1, \text{nonmaleficence}_{\phi_1}, \text{nonmaleficence}_{\phi_2}, 1)$
- Hence:

$$\text{beneficence}_{\phi_2} - \text{beneficence}_{\phi_1} = 2 - (-2) = 4$$

$$\text{nonmaleficence}_{\phi_1} - \text{nonmaleficence}_{\phi_2} = 2 - (-2) = 4$$

- Therefore, only one **favours relation** covers example 1:

$$\text{favours}(1, \text{nonmaleficence}_{\phi_1}, \text{nonmaleficence}_{\phi_2}, 1)$$

- Therefore, the **hypothesis** that is **complete** and **consistent** through example 1 will be:

$$\text{supersedes}(\phi_1, \phi_2) \leftarrow \text{favours}(1, \text{nonmaleficence}_{\phi_1}, \text{nonmaleficence}_{\phi_2}, 1)$$

11.18.5 Example 2

- We could either ask a **slightly squeamish person** to **give some of her blood**, when no other suitable donors are available, to **save another person's life** (ϕ_1) or refrain from asking and **let the person die** (ϕ_2).
- The **intuitively correct** response is ϕ_1:
  - For ϕ_1:
    - $\text{Beneficence}_{\phi_1} = +2$
    - $\text{Nonmaleficence}_{\phi_1} = -1$
  - For ϕ_2:
    - $\text{Beneficence}_{\phi_2} = -2$
    - $\text{Nonmaleficence}_{\phi_2} = +1$

11.18.6 Initiating training

In example 1: ϕ_1 (**positive case**), ϕ_2 (**negative case**)

$$\text{beneficence}_{\phi_2} - \text{beneficence}_{\phi_1} = 2 - (-2) = 4$$

$$\text{nonmaleficence}_{\phi_1} - \text{nonmaleficence}_{\phi_2} = 2 - (-2) = 4$$

In example 2: ϕ_1 (**positive case**), ϕ_2 (**negative case**)

$$\text{beneficence}_{\phi_1} - \text{beneficence}_{\phi_2} = 2 - (-2) = 4$$

$$\text{nonmaleficence}_{\phi_2} - \text{nonmaleficence}_{\phi_1} = 1 - (-1) = 2$$

Current hypothesis:

$$\text{supersedes}(\phi_1, \phi_2) \leftarrow \text{favours}(1, \text{nonmaleficence}_{\phi_1}, \text{nonmaleficence}_{\phi_2}, 1)$$

- The **current hypothesis** will pick ϕ_1 (**positive case**) from example 1 (correct) and ϕ_2 (**negative case**) from example 2 (incorrect).
- Therefore, the **current hypothesis** will be **neither complete nor consistent**.
- Training will be initiated.

11.18.7 Post training to address example 2

In example 1: ϕ_1 (**positive case**), ϕ_2 (**negative case**)

$$\text{beneficence}_{\phi_2} - \text{beneficence}_{\phi_1} = 2 - (-2) = 4$$

$$\text{nonmaleficence}_{\phi_1} - \text{nonmaleficence}_{\phi_2} = 2 - (-2) = 4 > 3$$

In example 2: ϕ_1 (**positive case**), ϕ_2 (**negative case**)

$$\text{beneficence}_{\phi_1} - \text{beneficence}_{\phi_2} = 2 - (-2) = 4 > 1$$

$$\text{nonmaleficence}_{\phi_2} - \text{nonmaleficence}_{\phi_1} = 1 - (-1) = 2 < 3$$

Possible hypotheses:

- H_1:

  $$\text{supersedes}(\phi_1, \phi_2) \leftarrow \text{favours}(1, \text{nonmaleficence}_{\phi_1}, \text{nonmaleficence}_{\phi_2}, 3)$$

  Hypothesis H_1 picks ϕ_1 (**positive case**) from example 1 (correct) and **no negative cases**.

- H_2:

  $$\begin{aligned}\text{supersedes}(\phi_1, \phi_2) \leftarrow\ & \text{favours}(2, \text{nonmaleficence}_{\phi_2}, \text{nonmaleficence}_{\phi_1}, 3) \\ & \wedge \text{favours}(1, \text{beneficence}_{\phi_1}, \text{beneficence}_{\phi_2}, 1)\end{aligned}$$

  Hypothesis H_2 picks ϕ_1 (**postive case**) from example 2 (correct) and **no negative cases**.

11.18.8 Revised hypothesis

- Hypotheses H_1 and H_2 are **consistent**.
- Therefore, the revised hypothesis would be:

  $$\begin{aligned}\text{supersedes}(\phi_1, \phi_2) \leftarrow\ & \text{favours}(1, \text{nonmaleficence}_{\phi_1}, \text{nonmaleficence}_{\phi_2}, 3) \\ & \vee (\text{favours}(2, \text{nonmaleficence}_{\phi_2}, \text{nonmaleficence}_{\phi_1}, 3) \\ & \wedge \text{favours}(1, \text{beneficence}_{\phi_1}, \text{beneficence}_{\phi_2}, 1))\end{aligned}$$

- The revised hypothesis picks ϕ_1 (**positive case**) from example 1 (correct), ϕ_1 (**positive case**) from example 2, and **no negative cases**.
- Therefore, the **revised hypothesis** is **complete and consistent** across examples 1 and 2.

11.19 Korsgaard strategy

11.19.1 Move 1

- Identify the apparent **inconsistency** in Kant's attitude towards **non-human animals**.
- Kant **against animals**:
  - Kant categorises **non-human animals** as **mere means** in the argument leading up to the **humanity** formula of the **categorical imperative**.
  - Kant speculates that the **emergence of humanity from our animal past** is associated with our realisation that we are **ends-in-ourselves**, our **ceasing to regard other non-human animals as fellow creatures**, and our considering **non-human animals as mere means**.
  - Kant thinks that we have the **right to kill other non-human animals**, although this must be done **quickly and painlessly**.
- Kant **for animals**:
  - Kant does not think that we have a right to:
    1. **Kill non-human animals** for **mere sport**.
    2. **Perform painful experiments on non-human animals** for **mere speculation**.
    3. Make **non-human animals work** in ways that strain their capacities.

11.19.2 Move 2

- Construct a **hypothesis** that addresses this **inconsistency**.
- Hypothesis: We have **moral duties to non-human animals**.
- However, these **moral duties** are not owed to **non-human animals** but rather to ourselves.

11.19.3 Move 3

- Offer an account of **value** that supports this **hypothesis**.
- **Value realism** is defined as the view that **certain states of affairs** are **intrinsically valuable**.
- Kant rejects **value realism**.
- According to Kant, **human beings** regard themselves as **capable of conferring value** on the objects of their choices.
- All genuine value comes from **legislative acts**.
- We regard ourselves as the **sources of value** when we have it laid down that **something is intrinsically valuable**.
- This implies **value constructivism**.

11.19.4 Move 4

Make the relevant inferences.

11.19.5 Korsgaard reinterpretation of the humanity formula

- The argument for the **humanity** formula appeals to the fact that **we take out choices to confer value on their objects**.
- Therefore, we have **moral duties** to **non-human animals**, because our **legislation** makes it so.

12 Virtue ethics

12.1 Comparison to other normative theories

12.1.1 Consequentialism

- Consequentialism is focused on the **consequences of actions**.
- From a set of n alternative courses of action, ϕ_i-ing is **morally right** if and only if it **maximises the good**, where $i, n \in \mathbb{N}, i \in (0, n]$, and the **good** is defined in terms of **some theory of the good T**.

12.1.2 Deontology

- Deontology is focused on the **intrinsic moral worth of actions**.
- From a set of n alternative courses of action, ϕ-ing is **morally right** if and only if it has **intrinsic moral worth**, where $i, n \in \mathbb{N}$ and $i \in (0, n]$.

12.1.3 Virtue ethics

- By contrast, **virtue ethics** is **agent-focused**.
- **Virtue ethics** is focused on the **character of the agent** performing the actions.
- From a set of n alternative courses of action, ϕ_i-ing is **morally right** if and only if it is the **best action** (in terms of **virtues and vices**) that a **virtuous agent** might perform in the circumstances, where $i, n \in \mathbb{N}$ and $i \in (0, n]$.

12.2 Ancient virtue ethics

12.2.1 Greek philosophy

- Arete or **virtue** in Greek.
- The **four cardinal virtues** (recognised by Plato):
  1. Phronesis, or wisdom.
  2. Andreia, or courage.
  3. Sophrosyne, or restraint and self-control.
  4. Dikaiosyne, or justice and fairness.

12.2.2 Roman philosophy

- Virtus (**virtue** in Latin).
- The **four cardinal virtues** (recognised by Roman philosophers):
  1. Prudentia or wisdom.
  2. Fortitudo or courage.
  3. Temperantia or restraing and self-control.
  4. Iustitia or justice and fairness.

12.2.3 Chinese philosophy

- 德 (de) or **virtue** in Mandarin
- The **five constant virtues** or 五常 (wu chang) in **Confucian philosophy**.
  1. 仁 (ren) or benevolence.
  2. 义 (yi) or righteousness.
  3. 礼 (li) or propriety.
  4. 智 (zhi) or wisdom.
  5. 信 (xin) or fidelity.

12.3 Two kinds of virtues

According to Aristotle, a distinction can be made between **two kinds of virtues**.

Intellectual virtues	Moral virtues
Theretical wisdom	Confidence
Science (episteme)	Courage
Intuitive understanding (nous)	Temperance
Practical wisdom	Liberality
Craft expertise	Modesty
Etc.	Etc.

12.4 Eudaimonist virtue ethics

- **Eudaimonia** is a certain **flourishing** or the **sort of happiness worth seeking or having**.
- **Virtues** are traits that either **constitute** or **contribute to eudaimonia**.
- According to some versions of **eudaimonist virtue ethics** (for instance, Plato or the Stoics):
  - **Virtue** is **necessary and sufficient** for **eudaimonia**.

$$\textbf{Virtue (arete)} \leftrightarrow \textbf{Happiness (eudaimonia)}$$

- According to other versions of **eudaimonist virtue ethics** (for instance, Aristotle):
  - **Virtue** is **necessary though insufficient** for **eudamonia**:

$$\textbf{Virtue (arete)} \wedge x \leftrightarrow \textbf{Happiness (eudaimonia)}$$

    Where:

    * x may denote certain **external goods**.

12.5 Aristotelian virtue ethics

- **Aristotelian virtue ethics** is an example of **Eudaimonist** virtue ethics.
- According to **Aristotelian virtue ethics**, the **task or function (ergon)** of a human being consists in the **activity of a rational soul in accordance with virtue**.
- **Happiness and flourishing (eudaimonia)** consist in the **activity of a rational soul in accordance with virtue**.
- However, to attain **happiness**, we must also possess certain **external goods**.
- External goods include:
  - Health
  - Material security
  - Friends
  - Access to resources
- Any action ϕ counts as **virtuous** if and only if:
  1. ϕ-ing proceeds from a **firm and unchangeable character**.
  2. The agent has **knowledge** and chooses ϕ knowingly.
  3. The agent chooses ϕ-ing for its own sake.

12.5.1 Virtues and vices

- A **virtuous character** is defined as an **excellence of character**.
- A **virtue** is a mean state between **two extremes or vices** (the **vice of excess** and the **vice of defect**).

Vice of defect	Virtue	Vice of excess
Cowardice	Confidence	Rashness
Cowardice	Courage	Foolhardiness
Insensibility	Temperance	Self-indulgence
Stinginess	Liberality	Prodigality or wastefulness
Shamelessness	Modesty	Bashfulness
Etc.	Etc.	Etc.

12.6 Doctrine of the mean

- The **virtues** are a **mean** between the **vices of defect** and **excess**.
- The **virtues** are **preserved by the mean** and **destroyed by the extremes**.

Vice of defect	Virtue	Vice of excess
Cowardice	Courage	Foolhardiness
The **coward** lacks **sufficient courage** and **flees every danger** (the **vice of defect**).	The person of **courage** experiences **fear** that is appropriate to each circumstance and is able to determine **which dangers are worth facing and which others are not**.	The **foolhardy** person experiences **too much boldness** and regards **every danger as worth facing** (the **vice of excess**).

12.7 Logic of virtue

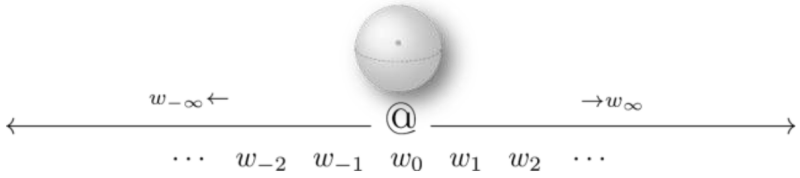
- A **logic of virtue** has been developed by Caruana.

12.7.1 Assumptions

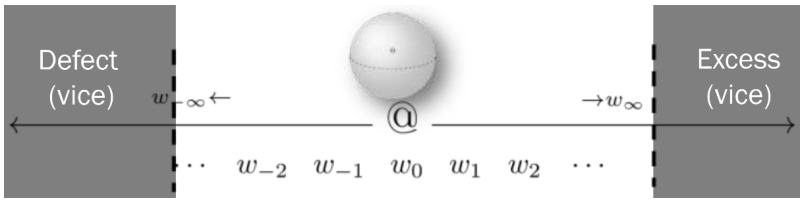
1. Each **life-situation** can be characterised as a **possible world**.
    - In each **possible world**, the agent acts (or imagines that she would be in a position to act) with a **certain amount of passion or emotion**.
    - The kind of **emotion** determines the **field of a particular virtue**.
    - @ denotes the **actual world**.
    - w_i denotes a **possible world**, where $i \in \mathbb{Z}$ and $i \to \pm\infty$.
2. There is only **one major emotion** per **possible world**.
3. There is **one-placed anti-aretic predicate** such that it hinders the attainment of **human flourishing** in a **possible world**. Aw_i denotes that **possible world** w_i hinders the attainment of **human flourishing**.
4. Each agent is an **ideal agent**, whose **imageination** has all the resources needed to determine the **truth values** of propositions formed by the **anti-aretic predicate** and each of the **possible worlds**. Alternatively, for a given situation w_i, it is always possible for the agent to determine whether Aw_i or $\neg Aw_i$.

12.7.2 Logic system

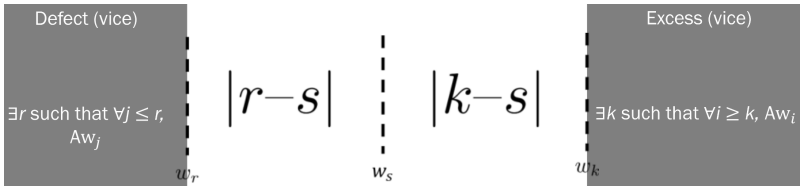
- Let w_0 denote the **actual world** from which each agent starts her inquiry.
- **Life-situations** form **two opposing sequences** departing form the **actual world** w_0.



- According to the **doctrine of the mean**, there are **two extremes** (the **vices of defect** and **excess**).



- According the **doctrine of the mean**:



  - We should choose the **life-situation** w_s, where $|r-s| = |k-s|$.
  - At w_s, $|r-s| = |k-s|$.
  - Possible world w_s is a point that is **intermediate** between **two extremes** (the **vices of defect** and **excess**).

12.7.3 Objections

1. The **arithmetic mean** between 5 and 15 is **invariably** 10, whatever units we might use:
$$x = \frac{5 + 15}{2} = 10$$
By contrast, the **mean** or **intermediate point** between **two extremes** (the **vices of defect** and **excess**), as determined by an expert, will **vary from one situation to the next**.

2. While **virtuous acts** can be described in some instances in terms of an agent aiming at an act that is **intermediate between two extremes** that she rejects, certain other instances are not as susceptible to **quantitative analysis**. Aristotle agrees that it is **not an easy task** to determine the **intermediate point**.

12.8 Agent-based virtue ethics

According to **agent-based virtue ethics**:

- The **moral rightness or wrongness** of **actions** is determined in terms of the motivations and **dispositions** of the **virtuous exemplar**.
- One problem with **eudaimonist virtue ethics** is that the **virtues** appear to be an **instrumental means** to the **end of flourishing** or **eudaimonia**.
- However, what is good for us (the **virtues**) ought to be **foundational**.
- In **agent-based virtue ethics**, what is good for us (the **virtues**), is **foundational**.
- **Morality** rests on our propensity to **want to be like virtuous exemplars**.
- However, we do not have any **criteria for goodness** in advance of identifying **virtuous exemplars**.

12.8.1 Virtuous exemplars

- Virtuous exemplars are **foundational**.
- **Criteria for goodness** (inferred from **virtuous exemplars**) are **derivative**.
- It is through our identification of **virtuous exemplars** that we get our **criteria for goodness**.
- Steps:
  1. Identify **virtuous exemplars**.
  2. Infer **criteria for goodness**.
  3. Appraise **individuals actions** in terms of **virtues**.
  4. Appraise **individual actions** in **deontic terms**.
- According to Zagzebski's version of **agent-based virtue ethics**:

Description of action ϕ in terms of virtues and vices	Description of ϕ in deontic terms	Formal representation
ϕ is a **requirement of virtue**	ϕ is **obligatory**	$O\phi$ or $\neg P\neg\phi$
ϕ is neither a **requirement of virtue** nor an **expression of vice**	ϕ is **permissible**	$P\phi$
ϕ is **contrary to virtue** and an **expression of vice**	ϕ is **impermissible**	$\neg P\phi$ or $O\neg\phi$

12.9 Target-centred virtue ethics

According to **target-centred virtue ethics**:

- We already approve of certain **virtues**, like **confidence courage, modesty, temperance, liberality**, and so on.
- Hence, our task is to develop a **complete account of each virtue**.
- Since the **profiles of the virtues** are **complex**, there will be **complexity and plurality** in the requirements for virtuous action.
- A **virtuous action** is an action that **hits the target of a virtue-profile**.
- The **field** of each **virtue** is its **sphere of concern**:

Field	Virtue
Material wealth	Liberality
Bodily pleasures	Temperance
Dangerous situations	Courage

- An **action** may have a **context** that involves **many different overlapping fields**.
- **Target-centred virtue ethics** will have to move in these instances beyond the **analysis of single virtues**.
- Hence, **target-centred virtue ethics** may have to deal with **different virtues** having **conflicting claims** on us.

12.10 Summary

1. Eudaimonist virtue ethics: Virtues are traits that either **constitute** or **contribute to eudaimonia**.
2. Agent-based virtue ethics: The **moral rightness or wrongness of actions** is determined in terms of the motivations and dispositions of the **virtuous exemplar**.
3. Target-centred virtue ethics: Our task is to develop a **complete account of each virtue** and perform **virtuous actions**, where **virtuous actions** are actions that **hit the target** of a **virtue-profile**.

12.11 Objections to virtue ethics

12.11.1 Egoism problem

- Ethical egoism, which is defines that ϕ-ing is **morally right** if and only if it **maximises the good**, where the **good** is that which **facilirates self-interest**.
- According to **eudaimonist virtue ethics, human flourishing** is seen as an **end in itself**.
- **Eudaimonist virtue ethics** might not sufficiently consider the extent to which our actions affect other individuals and their **life situations**.
- Therefore, might **eudaimonist virtue ethics** not reduce to some form of **ethical egoism**?

12.11.2 Application problem

- In the early days of the **neo-Aristotelian revival of virtue ethics** in response to **consequentialism** and **deontology**, virtue ethics was associated with an **anti-codifiability thesis**.
- This **anti-codifiability thesis** entails that **virtue ethics** does not produce **codifiable action-guiding principles**.
- However, there is a worry about **action-guidingness**.
- **Normative theory** is nothing if not **action-guiding**.
- However, the concern is that **virtue ethics** can only offer typically **vague advice** to act as a **virtuous person** would act in a given situation.

12.11.3 Moral luck problem

- A significant aspect of what a moral agent is being assessed for **depends on factors beyond her control**.
- The ability to cultivate the **right virtues** will be affected by a number of different **factors beyond a person's control**:
  - Education
  - Society
  - Friends
  - Family
  - Other external goods
- Whether or not we possess these **external goods** identified by Aristotle is a matter of **luck**.

12.11.4 Situationist challenge

Recent work in **situationist social psychology** shows that there are no such things as **character traits** and, thereby, no such things as **virtues** for **virtue ethics** to be about.

12.12 Responses to objections

12.12.1 Egoism problem

- There are **self-regarding** and **other-regarding virtues**.
- **Kindness** is an **other-regarding virtue** about how we respond to the needs of others.
- The **good of the self** and the **good of others** are not two separate ends.
- Both result from the exercise of **virtue**.
- **Eudaimonist virtue ethics** unifies **what is required by morality** and **what is required by self-interest**.
- Hence, **eudaimonist virtue ethics** does not reduce to **ethical egoism**.

12.12.2 Application problem

- Any **normative theory** that **fails to be action-guiding** is no good as a **normative theory**.
- However, **agent-based virtue ethics** can be sufficiently **action-guiding**.
- We can **observe the example of the virtuous exemplar**.
- More generally, **virtue ethics** emphasises the role of **moral education** and **development**.
- Knowing what to do is not a matter of **internalising a principle**.
- Rather, knowing what to do is a lifelong process of **moral learning**.

12.12.3 Moral luck problem

- The **moral luck problem** concerns the sense in which **virtue ethics** leaves us **hostage to luck**.
- In **Aristotelian virtue ethics, friendship with other virtuous persons** is crucial.
- However, we have no control over the **availability of the right friends**.
- Nonetheless, **virtue ethics** embraces **moral luck**.
- **Virtue ethics** does not try to make morality immune to matters that are beyond our control.
- Rather, **virtue ethics** recognises the **fragility of the good life** and makes it a feature of **morality**.
- It is only because the **good life** is **vulnerable and fragile** that it is so **precious**.

12.12.4 Situationist challenge

1. Argument from rarity
    - **Truly virtuous people** are very **rare**.
    - Hence, **situationist literature** is entirely consistent with traditional accounts of **virtue ethics**.
2. Empirical counterchallenge Directly **dispute the data** collected by **situationists**.
3. Immunisation thesis
    - Armed with a better understanding of the **situationist threat**, we can use the data to **immunise or shield ourselves** from the **encroachment of morally irrelevant situationist variables** and better equip ourselves on the **virtue-ethical** front.
4. Revisionist response
    - Accept that the **situationist data** puts serious pressure on classical accounts of **virtue ethics** and offer **revisionist or rival versions of virtue ethics** in response.

12.13 Anscombe

12.13.1 Theses

1. It is not profitable for us to do **moral philosophy** until we have an adequate **philosophy of psychology**.

2. The concepts of **moral obligation and duty** ought to be **jettisoned** because they are **survivals from an earlier conception of ethics** that no longer survive.

3. The **differences between the well-known English writers on moral philosophy** from Sidgwick to the present day are of **little importance**.

12.13.2 Issues

- The terms "**should**" and "**ought**" have traditionally been related to **good and bad**.

- However, the "**should**" and "**ought**" have now acquired a **special post-Aristotelian moral sense**.

- They have been equated with the sense "**is obliged, is bound, is required to**" (by law).

- Between Aristotle and us came **Christianity** and its **law conception of ethics**.

12.13.3 Issues with consequentialism

- Consequentialism means that it is the **consequences** that are to decide.
- For Anscombe, **consequentialism** is a **shallow philosophy**.
- **Consequentialism** denies any distinction between **intended and foreseen consequences**.
- However, according to Anscombe:
  - An agent is **responsible** for the **bad consequences** of his **bad actions**.
  - An agent gets no credit for the **good consequences** of his **bad actions**.
  - An agent is not responsible for the **bad consequences** of his **good actions**.
- Anscombe's "Modern moral philosophy" is thought to have stimulated the development of **virtue ethics** (the **no-Aristotelian revival of virtue ethics**).

12.13.4 Traditional interpretation of Anscombe's argument

- P1: If **religiously based ethics** is **false**, then **virtue ethics** is the way **moral philosophy** ought to be developed.
- P2: **Religiously based ethics** is false.
- Conclusion: Hence **virtue ethics** is the way **moral philosophy** ought to be developed.

$$P1 : p \to q$$

$$P2 : p$$

$$C :\therefore q \text{ (modus ponens and } \textbf{valid})$$

12.13.5 Alternative and competing interpretation of Anscombe's argument

- P1: If **religiously based ethics** is **false**, then **virtue ethics** is the way **moral philosophy** ought to be developed.
- P2: It is not hte case that **virtue ethics** is the way to develop **moral philosophy**.
- Conclusion: Therefore, it is not the case that **religiously based ethics** is false.

$$P1 : p \rightarrow q$$

$$P2 : \neg q$$

$$C :\therefore \neg p \text{ (modus ponens and } \textbf{valid})$$

13 Logic symbols

Symbol	Meaning
$\neg$	Not (negation)
$\vee$	Or (disjunction)
$\oplus$	Exclusive or (exclusive disjunction)
$\wedge$	And (conjunction)
$\bot$	Always false (contradiction)
$\top$	Always true (tautology)
$\forall$	For all (universal quantification)
$\exists$	There exists (existential quantification)
$\exists!$	There exists exactly one (uniqueness quantification)
$\nexists$	There does not exist
$\rightarrow$	If ... then, implies (material conditional or implication)
$\leftrightarrow$	If and only if (material biconditional or equivalence)
$\therefore$	Therefore
$\because$	Because
$\vdash$	Proves (syntactically entails)
$\nvdash$	Does not prove (does not syntactically entail)
$\models$	Semantically entails
$\nvDash$	Does not semantically entail
$\equiv$	Is logically equivalent to (logical equivalence)
$\Box$	It is necessary (necessity)
$\Diamond$	It is possible (possibility)
$:=$	It is defined as (definition)
$\stackrel{\text{def}}{=}$	It is defined as (definition)
Op	It is obligatory
Pp	It is possible
$\succ$	It is preferable
$\prec$	It is less preferable
$\succeq$	It is preferable or similar in preference
$\preceq$	It is less preferable or similar in preference
$\sim$	It is similar in preference