7CCSMEAI
Philosophy and Ethics of Artificial Intelligence

If you spot anything wrong, please let me know!
discord: @insertish
izzy · insrt.uk

Contents

1. Introduction

1.1. What is philosophy?

Philosophy is an activity undertaken to understand fundemental truths around oneself, the world we live in, and relationships to world and to each other. Literally means the love of wisdom.

Metaphysics is the study of the nature of reality, of what exists in the world, what it is like, and how it is ordered.

- Is there a God?
- What is truth?
- What is a person? What makes a person the same through time? What is the self?
- Is the world strictly composed of matter?
- Do people have minds? If so, how is the mind related to the body?
- Do people have free will?

Epistemology is the study of knowledge. It is primarily concerned with what we can know about the world and how we can know it.

- What is knowledge?
- Do we know anything at all?
- How do we know what we know is true?
- Can we be justified in claiming to know certain things?

Ethics concerns what we ought to do and what it would be best to do.

- What is good? What makes actions or people good?
- What is right? What makes actions right?
- Is morality objective or subjective?
- How should I treat others?

Logic is the arguments or reasons given for answers to questions.

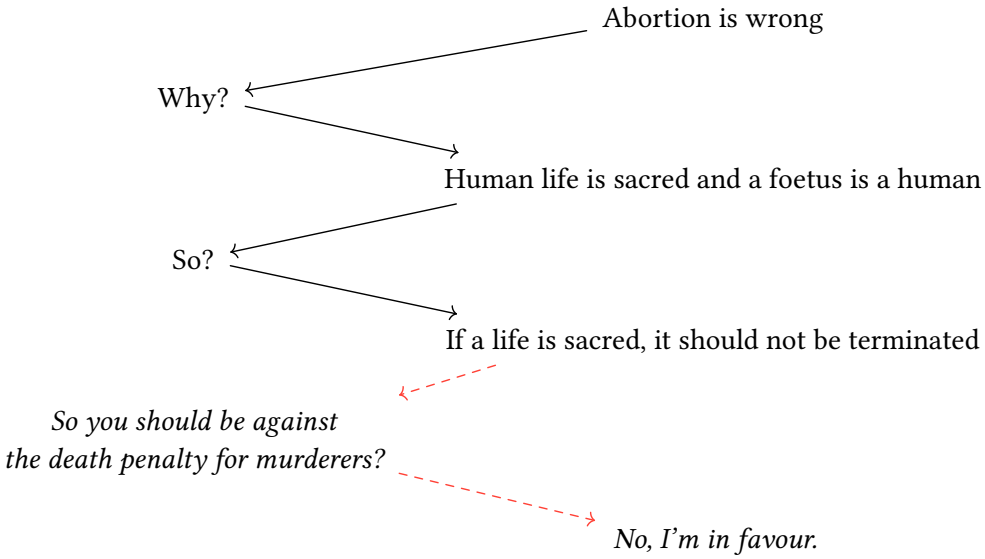
- What constitutes 'good' or 'bad' reasoning?
- How do we determine whether a given piece of reasoning is good or bad?

1.2. Impact of philosophy

> **Note**
>
> ### 1.2.1. Example 1: “I think therefore I am”
>
> Rene Descartes wrote this in 1637 to highlight the very act of thinking establishes an absolute proof of the existience of a self (the I) that is doing the thinking. Whereas everything else (perceived external world) could be an illusion conjured up by a demon to deceive us.
>
> We can only be certain of the existence of the self but we cannot be absolutely certain about anything else, including the physical world & body (could be a brain wired up to a machine). Therefore self/mind must be something fundamentally different from physical world.
>
> **Impact:** separation of mind & body historically permeated Western thought. For example, in Western medicine, the idea that one’s mental processes had any effect on physical state would’ve been dismissed in the past.

> **Note**
>
> **1.2.2. Example 2: "Free will, the self, responsibility and justice"**
> - Physics ultimately describes the nature and workings of the world incl. human behaviour.
> - All physical processes are *deterministic* - A is caused by B is caused by C and so forth.
> - Therefore free will is an illusion - that we have a 'genuine' choice rather than it being a choice determined by the underlying neural processes in the brain, influenced by inputs from the environment.
>
> **Impact:** so when we hold a criminal responsible for a crime, there is no *self* making a 'free' descision that we hold responsible, the crime was a consequence of a deterministic series of causes and effects.
>
> Therefore penalties for a crime should not be based on a desire to punish a self that is morally responsible. The penalty should be that:
>
> - it is incorporated into the chain of reasoning responsible for future decisions, such that decisions will not be made to reoffend (*rehabilitation*)
> - acts as a *deterrent* to others
> - *protects* society from the criminal and vice-versa

So when we apply these ideas to AIs making decisions in the future, and when things go wrong/ harms caused, to who should we assign moral responsibility and what actions should be taken to prevent harm in the future?

> **Note**
>
> **1.2.3. Example 3: "Logical Argument and Ethics"**
> Socrates used a method of questioning to expose weakness/contradiction in adversaries' arguments.
>
> 
>
>
> In the first part of the conversation, Socrates reveals Sara's line of reasoning/argument.
>
> - If *X is a human* then *X's life is sacred. A foetus is a human.* $\therefore$ *A foetus's life is sacred.*
> - If *X's life is sacred* then *X should not be terminated. A foetus's life is sacred.* $\therefore$ *A foetus should not be terminated.*

Then Socrates has revealed a contradiction, and weakened her argument. Given her stance against abortion, she should be against the death penalty however she is in favour.

- If *X is a human* then *X's life is sacred. A murderer is a human. ∴ A murderer's life is sacred.*
- If *X's life is sacred* then *X should not be terminated. A murderer's life is sacred.*
  ∴ A murderer should not be terminated.

1.3. Refining knowledge through refutation

Knowledge often evolves through refutation (e.g. scientific knowledge) so that seeing a bird flapping its wings but failing to fly: we may change the rule, such as in the case of penguins.

We make the change so there is no contradiction:
- If *X is a bird* then *X flies.*
- If *X is a bird* and *X is not a penguin* then *X flies.*

Note

1.3.1. Example 3 (cont.): "Socrates v. Sara"

Sara can adjust her argument to say that if *X is a human* and *X does not take the life of another human* then *X's life is sacred.*

- If *X is a human* and *X does not take the life of another human* then *X's life is sacred.*
  A foetus is a human.
  A foetus does not take the life of another human.
  ∴ A foetus's life is sacred.
- If *X's life is sacred* then *X should not be terminated. A foetus's life is sacred.*
  ∴ A foetus should not be terminated.

In this case, a murderer is a human but **does** take the life of another human, so Sara can use this line of reasoning.

However, we can further challenge this reasoning by asking about whether an executioner's life should be sacred, after all they are mandated by the state to take a life.

2. Intelligence

Intelligence can be described as the ability to 'accomplish complex goals', the ability to 'acquire and apply knowledge and skills', and 'the faculty of understanding; intellect' (capacity to understand).

The first definition (from Life 3.0) subsumes the other definitions, e.g. capacity for logic, 'understanding', self-awareness, learning.

> **Note**
>
> Lets suppose the first definition, how can we say one being is more intelligent than another? Is a Go or Chess playing computer more intelligent? What sort of intelligence ordering do we obtain from "X is more intelligent than Y if all goals accomplished by Y can be accomplished at least as well as X, and X can accomplish one goal better than Y".

Anthropomorphic perspective rates certain tasks as more intelligent than others, such as complex multiplication as opposed to recognising a friend in a photo. **Moravec's Paradox** refers to the phenomenon where tasks that are easy for humans to perform are difficult for machines (motor/social skills) and vice-versa (mathematical calculations/large-scale analysis).

From an AI perspective, intelligence is ultimately about information and computation and not dependent on underlying 'hardware'. We say it is **substrate neutral**, and so there is no reason that machines cannot be as (or more) intelligent as humans.

2.1. Information

Information storage devices have components arranged in a way that is related to the state that the information is about.

Fundamental physical property of information is that they can be in many different long lasting (stable) states that take more energy to displace than provided by random disturbances.

The simplest memory device has two stable states which encode a binary digit (0 or 1). This can be physically embodied in many different ways (magnetisation, positions of electrons, strong/weak laser beams, etc).

We say that information is **substrate neutral/independent**.

2.2. Computation

Computation transforms one information state into another, by implementing a function. Highly complex functions imply accomplishment of highly complex goals.

Turing completeness means something is equivalent to a Turing machine which is computationally universal in the sense it can do anything any other computer can do - given enough time and resources.

> **Note**
>
> NAND gates are Turing complete because we can implement any well-defined function by connecting NAND gates together.

A **universally intelligence machine** is one that given enough time and resources, it can accomplish any goal as well as any other intelligent entity.

We say that computation is **substrate neutral/independent**.

2.3. Building Artificial Intelligence

AI researchers are said to build systems that fall into one of the following quadrants:

	Human-based	**Ideal Rationality**
Reasoning-based	Systems that think like humans	Systems that think rationally
Behaviour-based	Systems that act like humans	Systems that act rationally

> **Note**
>
> **Russell and Norvig** (*Artificial Intelligence: A Modern Approach*) fall within the acting rationally camp:
> - "what is AI?" question is recast as "what is intelligence?" and then identified as acting rationally
> - **AI is the field devoted to building intelligent agents** which are functions taking as input, percepts from external environment, and producing behaviour (actions) on basis of these precepts.

A **perfectly rational agent** is one that models the environment E considering actions $a_1, ..., a_n$ that result in changes to states of the world, hence new environments $E_1, ..., E_n$, and then multiplies probability each a_i will result in E_i by the value/worth/utility of E_i to then act to bring about E_i with highest expected utility.

Usually it is not possible to build perfectly rational agents.

Calculatively rational programs are that if executed infinitely fast, they would result in perfectly rational behaviour.

Bounded Optimality is given a machine M (with associated time/space constraints) the goal of finding the optimal program given environment E resulting in agent acting to maximise expected utility.

AIXI is a mathematically rigorous framework[1][2] for defining optimal intelligence and provides a definition of intelligence that generalises the Russell and Norvig definition.

It defines that intelligence as a measure of an agent's ability to achieve goals in a *wide range of environments*. Represented by a set of meaningful equations.

Optimal (maximally) intelligent agent according to these equations is not Turing complete - challenge is to integrate resource bounds that approximate theoretical ideal.

2.4. Types of Intelligence

Narrow AI are purpose-built algorithms for specific tasks/goals which often outperform human capabilities.

Artificial General Intelligence (AGI) is the ability to accomplish any goal at least as well as humans. Possessing common sense and an ability to learn, reason, and plan, to meet complex information processing challenges across a wide range of natural and abstract domains.

[1] https://plato.stanford.edu/entries/artificial-intelligence/aixi.html
[2] https://jan.leike.name/AIXI.html

2.5. Moravec's Landscape of Human Competence

Moravec imagined a 'landscape of human competence' where advancing computer performance is like water slowly flooding the landscape.

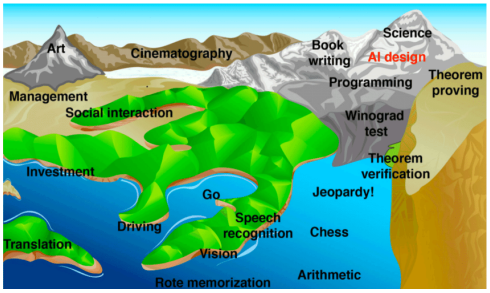


Figure 1: Moravec's Landscape

2.6. Path to AGI/Superintelligence

One interpretation of the **singularity** is the point at machines can improve machines faster than humans can improve them. In Moravec's landscape of human competence, the critical sea level relates to when machines can perform AI design.

> **Note**
>
> Most futurists settle on 20 years before the advent of AGI, attention grabbing and relevant, yet far enough to imagine breakthroughs.

2.6.1. Common issues with early AI

Early AI adopted logic/symbolic paradigms - rules applied to data.
But common issues include:

- Problems handling uncertainity
- *brittleness, symbol grounding problem*
  - brittleness: small damage to logical knowledge base implies total crash
  - **symbol grounding problem**: humans 'hard coding' initial beliefs

2.6.2. Recent work on AI and the future

More recent work involves:

- Focus on high level symbol manipulation, neural networks that degrade gracefully due to small damage
- Learning from experience (to try to solve symbol grounding problem) and finding natural ways of generalising from examples
- Back propagation algorithms providing ability to learn a much wider range of functions
- Advances in hardware and processing power
- Data fuelled explosion in use of NNs and ML

Expert community survey results predict 10% chance that AGI is reached by 2022, 50% by 2040, and 90% by 2075.

Once we have AGI, superintelligence is argued to be a small step away.
Superintelligence is any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest.

The most popular interpretation of the **technological singularity** is the hypothetical future point in time at which technological growth becomes uncontrollable and irreversible, resulting in unimaginable changes to human civilisation.

2.7. The Imitation Game (The Turing Test)

How will we know when we achieve AGI?

- Human level performance in common sense reasoning (e.g. frame problem) and natural language understanding are considered AI complete problems: difficulty of solving these is equivalent to difficulty of building AGI.
- Alan Turing's 1950 paper "Computing Machinery and Intelligence" takes the approach of 'linguistic indistinguishability'
  He begins the paper with claim "I propose to consider 'Can machines think?'"
  He rejects traditional approach of defining machine and intelligence and replaces question with a new one: 'Can machines do what we (as thinking entities) can do?'
  Turing argues that this question draws a 'fairly sharp line between physical and intellectual capacities of a man'.

The Imitation Game (The Turing Test)

- Suppose an interrogator is in a room separated from a person and a machine.
- Object of the game is for the interrogator to determine which of the two is the person/machine.
- Interrogator is allowed to put questions to the person and machine.
- Object of the machine is to try to cause the interrogator to mistakenly conclude that the machine is the other person.
- Object of the other person is to try to help the interrogator to correctly identify the machine.

Given the scenario that a woman and a computer are in sealed rooms, and a human judge doesn't know which of the two rooms contains the contestant: if, on the strength of the returned answers, the judge can do no better than 50/50 when delivering a verdict as to which room houses which player, we say that the computer in question has passed the turing test.

Computer demonstrates that its responses are indistinguishable from a human, at least linguistically, so we say that passin gin this sense operationalises linguistic indistinguishability.

> **Note**
>
> Turing anticipates importance of ML as means of bootstrapping a simpler system to human intelligence, i.e. instead of trying to produce a programme to simulate the adult mind, why not produce one to simulate the child and then subject appropriate course of education.

> **Note**
>
> Rene Descartes also prefigured this sort of thinking:
> - 'If there were machines which bore a resemblance to our body and imitated our actions as far as it was morally possible to do so, we should always have two very certain tests by which to recognise that, for all that, they were not real men. The first is, that they could never use speech or other signs as we do when placing our thoughts on record for the benefit of others. [..] But it never happens that it arranges its speech in various ways, in order to reply appropriately to everything that may be said in its presence, as even the lowest type of man can do'
> - Descartes thought it was impossible a mere machine could produce different arrangements of words so to give a meaningful enough answer.

2.8. Analysis of the Turing Test

Empirical questions about if it's possible:

- Is it true we will soon have computers that play the game so well an average interrogator has no more than 70% chance of correct identification after 5 minutes?
  - Claims in 2014, program Eugene Goostman fooled 33% of judges in Turing Test competition.
  - Similar results in other one-off competitions.
  - Not reliabily reproducible results, sizes of trials small.
  - No strong grounds to support this.

2.8.1. Conceptual questions about what the Turing Test means

Conceptual questions about what TT means:

- One can interpret the TT as providing logically sufficient conditions for determining presence of intelligence. As in, $\text{PassTT} \rightarrow \text{Intelligence}$ or $\text{PassTT} \subseteq \text{Intelligence}$. Passing the test means intelligence is presence, therefore there may be intelligent beings that do not pass the TT.
- One can interpret the TT as providing logically necessary conditions for intelligence. As in, $\text{Intelligence} \rightarrow \text{PassTT}$ or $\text{Intelligence} \subseteq \text{PassTT}$. If something is intelligent, it must pass the TT.

2.8.1.1. Arguments about what TT means

1. *Logically sufficient and necessary conditions*: $\text{PassTT} \leftrightarrow \text{Intelligence}$
   Claimed by very few people, but some objections to TT only make sense assuming logically necessary interpretation which implies if something fails TT then it cannot be intelligence.
   chauvinistic objection: intelligent creatures may fail TT because they do not share out way of life (pragmatic conventions that govern the languages that they speak are very different from the pragmatic conventinos that govern human language)

   Note: this objection is only valid because Turing asks 'can Machines think?' as opposed to 'can a machine exhibit intelligent behaviour equivalent to, or indistinguishable from, that of a human?
2. *Only logically sufficient conditions*: $\text{PassTT} \rightarrow \text{Intelligence}$
   i.e. logically impossible if passing TT, then not intelligence
   Objections to this and other behavioural tests is a being whose behaviour was produced by brute force, e.g. consider Blockhead, a creature who looks just like a human being, but is controlled by a look-up tree that contains a programmed response for every discriminable input at each stage in the creature's life. So it can use the tree to find appropraite linguisitc response to every possible question.

   Assuming:
   - Blockhead is logically possible
   - since Blockhead uses brute force method of look-up, Blockhead is not intelligence

   Then Blockhead is an objection (counter-example) to *only logically sufficient conditions.*
   But we can deny logical possibility by (controversially) denying that:

   $$\text{conceivability} \rightarrow \text{logical possibility}$$

   Or insist that Blockhead is intelligent: as it is processing information together with its behaviour

3. *TT provides (more or less strong) probabilistic support for attribution of intelligence*
   Turing thought that passing TT would provide probabilistic support for hypothesis of intelligence:
   - prediction that Turing makes itself is probabilistic: 50 years, 70% right, five minutes each
   - probabilistic nature of prediction gives good reason to think TT itself is probabilistic
     a given level of success should produce a specifiable level of increase in confidence that participant is intelligent

   Since TT does not correlate success with increase in confidence, TT is greatly underspecified. Variables such as questioning period, skills and expertise of interrogator, skills and expertise of human player, etc are not specified.
4. *The TT is too hard*
   Because nothing without a human cognitive substrate could pass the test.
5. *The TT is too narrow*
   - Can be argued success might come for reasons other than possession of intelligence, but rather example of things that intelligent beings can do.
   - Success in TT implies cognitive competency in memory, perception, maths, game playing, understanding politics, etc
   - Ability to pass TT, esp. on repeated runs, implies ability to solve complex goals in wide range of environments
6. *The TT is too easy*
   Some argue a more demanding test (e.g. Lovelace Test) which judges AI based on ability to create/form an original idea.

2.9. Arguments against possibility of AGI

Philosopher John Searle uses a similar thought experiment to Blockhead to argue against Turing's claim that an appropriately programmed computer could think.

The **Chinese Room** attempts to show that a computer can only do manipulation of symbols, but has no true understanding which is one of the hallmarks of intelligence, so a computer cannot be intelligent.

Imagine a native English speaker who knows no Chinese locked in a room full of boxes of Chinese symbols (a data base) together with a book of instructions for manipulating the symbols (the program).

Imagine that people outside the room send in other Chinese symbols which, unknown to the person in the room, are questions in Chinese (the input). And imagine that by following the instructions in the program the man in the room is able to pass out Chinese symbols which are correct answers to the questions (the output).

The program enables the person in the room to pass the Turing Test for understanding Chinese but he does not understand a word of Chinese.

1. Symbol manipulating program P is a kind of digital computation that is possible in some remotely conveivable world
2. P does not possess understanding and hence is not intelligent
3. Digital computer C will not differ from P in any sense that makes it more intelligent than P.
4. $\therefore$ if P is not intelligent, then any actual digital computer C cannot be intelligent

Building off the Chinese Room argument:

- Given the assumption P does not possess understanding, hence is not intelligent: narrow conclusion programming a computer may make it appear to understand language but it does not possess real understanding. *Hence TT is inadequate test for intelligence.*
  Computers merely use syntactic rules to manipulate strings, but have no understanding of meaning/semantics
- Broader conclusion is that human minds are not like computers — simply information processing systems — as humans have real understanding.
  Although, our definition for intelligence — a system achieving complex goals in a wide variety of environments — does not imply a requirement for understanding. So can we consider it an argument against TT or as a test for intelligence/possibility of AGI? (only if assumption 2 holds)

2.9.1. Replies to Chinese Room Argument

- *Impossibility Reply*: challenges argument's assumption that the symbol manipulating program P is possible given laws of physics, hence argument and its claim is not valid. Philosopher Danniel Dennet argues P is not a nomic possibility given laws of physics; such a program to simulate the Chinese Turing test would likely be too long to run by hand or even digitally, and probably take many lifetimes before any Chinese came out.
- *System's Reply*: concede man in the room doesn't understand Chinese, but instead argue that the man is part of a larger system that includes a huge database, memory containing intermediate states, and the instructions. Given that, the system as a whole understands Chinese.
- *Virtual Mind (VM) Reply*: concede man in the room doesn't understand Chinese. State the whole system may create a virtual entity, distinct from system and its parts, that understands Chinese.
- *Robot Reply*: concede Searle is right that the whole system in room cannot understand langauge and know the meaning of words, but if let loose in the world (e.g. as a robot), it would come to understand meaning by interacting with the world and others.

  Someone in the room, or the room alone, cannot understand Chinese. But if you let the outside world have some impact on the room, meaning or 'semantics' hence understanding might begin to develop. But this concedes thinking cannot be just symbol manipulation?

3. Consciousness

Consciousness is subjective experience, something is conscious if there is something that is like to be that something, it can be described as 'something we're all familiar with, that goes away every night in deep sleep, and comes back in the morning, or whenever we dream. It encompasses all our subjective feelings and experiences. It's the only thing directly and immediately known to us and mediates our knowledge of the external world.', — Giulio Tononi (neuroscientist), '- and yet it is one of the most profound and seemingly insoluble mysteries of the world'.

Review of The Chinese Room as it relates to consciousness

The Chinese room argument assumes, intelligence → understanding, but as discussed previously, the question remains whether 'understanding', in the sense there is e.g. an explicit conscious awareness/apphrehension of relationship between Chinese symbols and meaning of said symbols, necessary for intelligence?

- i.e. if words can be used in all the ways that communicate the right kind of information for achieving complex goals, why do we need the extra internal apphrehension of what they mean?
- If we consider modern NLP models that are based on learnt context (sentences) in which words appear: the meaning of words are the totality of contexts in which they are used; meaning of a word is all the ways the word can be used.

⟶ so then is the Chinese room really only an argument against the posibility machines can have conscious experience — what it feels like to understand something?

To conclude:

- Chinese room is an argument against AGI if one accepts (a) understanding is necessary for intelligence and (b) understanding must include inner awareness or feeling that accompanies understanding.
- One may deny (a) as discussed in last section.
- One may deny (b) as illustrated by modern NLPs.
  The only way to recognise some entity understands something is by observing it can do everything implied by such an understanding.
  There may not be anything going on in the 'head' of a machine that is the *feeling* (conscious attendent) of understanding, but machine could be seen to take advantage of what it is to understand something in all the ways that one might imagine it is to take advantage of what it is to understand something.

Searle actually restated the conclusion of his argument 30 years later (2010):

> I demonstrated years ago with the so-called Chinese Room Argument that the implementation of the computer program is not itself sufficient for consciousness or intentionality.

Implication of this revised interpretation:

- Computer may behave (by virtue of replies it gives) as if there were real understanding when in fact there is none.
- But if a machine can do everything/behave as if there were real understanding, can we not say that it has 'real' understanding?

Searle argues against **functionalism**: the view that states/processes count as being of a given mental or conscious type because of the functional role they play.

If it looks like, walks like, and quacks like a duck, it is a duck.
If it behaves as if it has real understanding, then it has real understanding.

3.1. Why is Consciousness Important for AI?

Key questions

1. Is consciousness necessary for being intelligent?
   Definitions of Intelligence, AGI (ability to acheieve any goal at least as well as humans), and Superintelligence (intellect greatly exceeds cognitive performance of humans) do not require consciousness to be present.
   Doesn't necessarily mean consciousness is not needed for intelligence.
2. Typical definitions of consciousness do not exclude possibility of artificial consciousness.
   If future AI is conscious, they may experience pain/suffering and we should treat them accordingly.
3. If future AI conscious to more or less same extent as humans, this may have profound cultural, social, political, ethical implications, changing nature of human society.
   What if humans treat future AI as if were conscious irrespective of consciousness? What impact does this have on us?

Key explanatory problems

- the *easy* problem: how does the brain work in all of its functions, in all its detail. How does it interpret and respond to sensory inputs, how can it report on its internal state using language, how does the brain support all behaviours?
  These are very difficult problems to solve, but are problems of intelligence, they do not require solving 'what it feels like / subjective experience'.
- the *hard* problem: how does physical matter give rise to feelings/experience — what does it seem like from the inside?
  When you drive you experience colour, sound, emotion, feeling of self, do self-driving cars experience anything at all?
  In both cases, information is input from sensors, processed, and output as motor commands. We may ask, why does one arrangement of particles result in an experience but not another?

  Everything about life can be defined in terms of relationships between material parts, from metabolism to reproduction to perception — all explained in mechanistic terms/physics. The execption is consciousness, science cannot account for these existence of these properties.

3.2. Theories of Consciousness

Broadly speaking, there are two kinds of theories of consciousness in philosophy:

- **Dualist theories**
  - **substance dualism** is a dualist theory that states both physical and non-physical (conscious) substances exist — 'I think therefore I am'
  - **property dualism** is a dualist theory that states two distinct and disjoint kinds of properties — physical ones and conscious ones
- **Physicalist theories** propose everything is physical and property of consciousness in principle can be explained in terms of the physical
- Also, **neutral monism** proposes fundamental nature of reality is neutral, neither mental nor material

Thought experiment against Physicalism, and for Property Dualism

A **philosophical zombie** is an exact physical duplicate of a person that behaves in exactly the same way as a person - such as you - but lives in a possible world that is indistinguishable from the actual world except that such a zombie does not have an experiential consciousness.

Consider the argument:

- Zombies are conceivable
- If conceivable then possible - all same properties of actual world but not experiential property of consciousness
  note: notion of conceivability $\rightarrow$ possibility is controversial
- Now if consciousness were a physical property X, it would be impossible for a creature to have the same physical properties as you but not have a conscious experience Y. But such a creature is possible $\neg Y$, therefore consciousness is not a kind of physical property $\neg X$.
- Therefore property dualism is true — conscious properties are neither identical nor explainable in terms of (reducible to) physical properties.

Is such a zombie world even conceivable, and logically, nomically possible?

- More we understand about consciousness & brain processes underlying experiential consciousness -> less conceivable a zombie is.
- You can imagine a plane flying backwards, but the more you know about it, the harder it is to. e.g. their structure, aerodynamics, …

Even rejecting the zombie argument and assuming its not conceivable, there is still an explanatory gap — the *hard* problem — of explaining subjective feelings in terms of relations between material parts.

3.3. What is a Conscious State

The **notion of a conscious state** has distinct but interrelated meanings:

1. Simply a mental state one is aware of being in — to have conscious desire for something is not only to have such desire but be aware you have the desire.
   In principle, an AI could have such states of meta-awareness.
2. Involves qualitative/experiential properties called *qualia*: raw sensory feels, e.g. taste of wine, tacile feel of cheese; **qualia** are phenomenal properties, the phenomenalstructure of consciousness refers to overall structure of experience - not just sensory qualia but also much of the spatial, temporal, and conceptual organisation of our experience of the world and of ourselves as agents in it.
   Could AI have qualia? Encounters the *hard* problem again.
3. A state might be conscious in the sense that it relates to, is available to, and interacts with other states. This is **access consciousness**. A visual state's being conscious is not about whether there are associated qualia but whether or not it and the visual information it carries is available for use and guidance by the organism.
   In principle, an AI could have such states.

3.4. The Problems of Consciousness

1. *what is consciousness? what are its features?* how does one give a description of qualia and the structure of phenomenological experience more generally? is it even possible to give a fully descriptive account of phenomenological experience from the outside? could there be inverted qualia — how do I know what 'red' you experience is the same as the 'red' I experience?
2. *how can consciousness exist?* returns us back to the *hard* problem, can we explain/understand how non-conscious items could cause/give rise to consciousness? normally one analyses macro-properties in terms of functions and then show how micro-structures obeying physical laws are enough to guarantee satisfaction of relevant functional conditions.
   Micro-properties of collections of H2O molecules at 20c suffice to satisfy conditions for liquidity of the water they compose. However, we have no satisfactory explanation of how consciousness is produced, **explanatory gap**.

   - Some argue human cognition limits us and we will never be able to bridge the gap, much like facts about democracy are *cognitively closed* to pigs.
   - Some argue the gap cannot in principle be closed, by any congitive agents, and argue this then means physicalism cannot be true.

   There is no way in which consciousness could be intelligbly explained as arising from physical, we can easily conclude it doesn't — that dualism is true — and if dualism is true, then this **suggests conscious AI may not be possible**.
3. why does consciousness exist?
   1. *can consciousness in principle play a causal role in behaviour?* if our behaviours are ultimately determined by physical process and our conscious experience is non-physical, then how can conscious experiences play a role in causal chain of neuronal/chemical events that result in behaviour? some argue it is **epiphenomenal** — does not play this role. indeed some features such as self-awareness and awareness of mental states when making decisions may just been seen as ways of reporting to ourselves of behaviours that are already underway.
   2. *if it does play a causal role, what is it?* what is its adaptive value from a evolution point of view? (present function, if it has one, might not be the same as which it evolved for) the brain's most basic function is to keep themselves and body alive — **interoception** — the feelings one is aware of in one's own body, emotions, mood, hunger, selfhood. these are arguably the brain's way of monitoring the body for signals indicating need to act to ensure survival; some more 'complex' emotions/feelings evolved for social reasons such as disgust, shame, etc (social brain hypothesis)

3.5. Roles that Consciousness Play

Suppose *epiphenomenalism* is false, the conscious does play a role. What sort of effects do features of consciousness have and what difference does it make? NB. does it enable intelligent behaviour and whether machines would benefit from consciousness

1. *self-/meta-awareness*: increases flexibility and sophistication of control (compared with unconscious processes which are by contrast efficient/speedy) which is important when dealing with novel situations and previously un-encountered problems — considered a hallmark of intelligence

   This also allows social coordination, being aware of our own beliefs, goals, desires, perceptinos enables us to have a better understanding of others' mental states. Mutually shared knowledge of others' minds enables interaction, cooperation, and communication in more advanced/adaptive ways.

   Arguably, could be implemented in a machine, if awareness is understood as having information about informational states without concerning ourselves of phenomenological character of 'meta' mental states.

2. *unified and densely integrated representations of reality*: perception of objects independently existing in time and space relies on integration of information from various sensory channels as well as background knowledge & memory. Not all sensory information necessarily needs to be experienced to affect behaviour but it can drive more sophisticated and flexible responses. Although, phenomenal experience may not be *necessary* for more unified integrated representations.

3.6. Information Theoretic Theories of Consciousness

What is the idea behind information theoretic theories of consciousness?

- Many notions of conscious states and roles of consciousness are described in informational terms: information about information (meta-awareness), integration of information, making information widely avaiable
- Suggests 'purely informational' theories of consciousness that do not necessarily concern themselves with solving the *hard* problem of phenomenological experience.
- Indeed recent theories try to characterise or 'explain' consciousness in terms of storage, processing, integration, and availability of information which suggest machine consciousness is possible.

The **Global Workspace Theory (GWT)** describes consciousness in terms of competition among processors and their inputs (sensory data, ideas, …) and outputs 'broadcasting' information for widespread access and use.

Particular processed outputs that 'win out' over other processed outputs then become available to global wrokspace and so accessible and influential with respect to other contents/prcoessors. At the same time, recurrent support (feedback) from workspace and from other contents influenced by what is broadcasted, strengthens the broadcasted content.

Availability and recurrent strengthening of broadcast information makes it conscious in the *access consciousness* sense.

3.6.1. Integrated Information Theory

The **Integrated Information Theory (IIT)** says consciousness is linked to the degree to which information is integrated, which in turn can be represented by precise mathematical quantity, ϕ. Parts of the human brain supporting consciousness have very high ϕ and therefore highly conscious, whereas systems with low ϕ have small amount of consciousness.

In principle, we could have a consciousness meter that measures ϕ and tells us the extent to which an organism is conscious.

IIT implies **panpsychism** — a variant of (property) dualism that says all the constituents of reality have conscious, or at least proto-conscious, properties distinct from whatever physical properties they may have.

The IIT's criteria for consciousness:

- *system has information about itself*: information about causal powers specifically, how much we can know about previous and next state of the system by looking at the state of the system now; current state of typical human brain can tell you a lot about what brain looked like a moment ago and what it will look like in the next moment
- *integration of information*: measures (ϕ) how much information of the system depends on interconnections between the system's parts, i.e. if we cut the system in half, how much information is lost?

  Consider cutting a book in half, you can still read the one half of a page then the other half, and it still conveys the same information. A highly interconnected brain cut in half would lose prior and subsequent state information if the information is highly integrated. This is a key difference between brains and computers — computers can have as much info as the brain and themselves — yet today's computers consist of transistors connected to a few others, that are feed-forward and not recurrent, so have very low ϕ.
- *maximality of integration*: conscious system must be a maximum of integrated information: it must have more integrated information that any of its parts and any bigger system of which it itself is a part of.

Implications of IIT for AI

- *rejects functionalist account of consciousness*: we reject 'if a system S1 produces same input/output as S2 that is conscious, then S1 is conscious, whether matter is the same is irrelevant' because of computers' spare interconnectivity (transistors connected to a few others) — feed-forward and not recurrent as discussed before, so have very low integration (ϕ)

  In "Ascribing Consciousness to Artificial Intelligence"[3], a scenario is put forwards to imagine what would happen if we create a digital simulation of the brain where in we graudally replace neurons with functionally equivalents and simulated all connections — would this simulation really gradually lose consciousness? No principle prevents computers in future being built with sufficient integration.
- rejecting IIT's insistence that digital computers cannot be conscious (accepting possibility of sufficient integration), arguably consciousness is *substrate independent*; so information theoretic view of consciousness combined with substrate independence, could suggest AI could be conscious

[3] arXiv, published April 2015, https://arxiv.org/abs/1504.05696

3.7. Conscious Machines

How would we recognise that an AI is conscious?

- it acts/behaves like it is conscious (functionalism)
- high ϕ or other information theoretic measure
- right kind of architecture

But we could never, even in principle, definitely say a machine is conscious since we cannot in principle experience the world from the inside in the same way a machine would, much like we can't in principle experience the world the same way other animals do.

There may have been evolutionary rationale for origins of consciousness. What would account for origins of machine consciousness? Would it simply emerge as machines become more intelligent and sophisticated, processing information in more and more complex ways? Does it even ultimately matter?

4. Reasoning and Communication

4.1. Logical (Symbolic) Approaches to AI

Early logic-based approaches to AI involved reasoning with symbolic formulae representing beliefs, desires, goals enabling reasoning about the way the world is (epistemic reasoning) and about what one should do (practical reasoning). Where the description of the world is in some logical languages (analogous to use of natural language to describe BDI) and inference ruels are applied to derive new facts about the world and appropriate actions to achieve goals.

4.1.1. Formalisation of Mathematics

A mathematical crisis at the beginning of 20th century → regarding foundations of mathematics ← triggered by discovery of various paradoxes. This motivated the development of a research program to try to prove all the truths of mathematics from given sets of axioms and then show that these systems are complete (all true statements are provable) and consistent. This program continued until Gödel's incompleteness theorems[4] showed this was impossible.

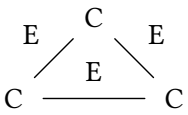
The research worked generally as follows:

- Start with a set of self-evidently true axioms (e.g. $\alpha \vee \neg\alpha$) and self-evident logical inference rules (e.g. $\frac{\alpha \vee \beta, \neg\alpha}{\beta}$).
- The truth of the axioms and applying inference rules until complex statement in question is deduced.
- If anything that can be inferred from the axioms is a true statement about the model, then the proof system is **sound**. If any true statement about the model can be inferred from the axioms then the proof system is **complete**.

Example: Proving Euclidean geometry from basic self-evident axioms

Euclid proved all truths of Euclidean geometry from 5 basic self-evident axioms/truths of 2D planes.

Take a simple example: "model/world is a triangle and we specify a language of symbols E,C representing set of edges and corners of the triangle"



The Axiomatic system (self-evident truths):

1. Every member of C is contained in exactly two members of E
2. Every member of E contains exactly two members of C
3. Every two members of E share exactly one member of C

One can state these axioms using symbols and logical connectives and quantifiers, then use rules of logical inference to infer/derive that every three members of E contain exactly three members of C.

Rules of logical inference compromise a proof system (syntax) and the triangle is the model (semantics).

[4] https://en.wikipedia.org/wiki/G%C3%B6del%27s_incompleteness_theorems

Good old fashioned AI (GOFAI) is a symbolic AI that adopted the methodology:

- Given a model W of the world, designer encodes axiomatic truths Δ_0 about W.

$$W \vDash \Delta_0$$

- in the logical knowledge base KB_{Δ_0} of agent (the agent's beliefs) and defines a (hopefully) sound and complete proof system (PS) for inferring more complex truths:

$$\mathrm{KB}_{\Delta_0} \underset{\mathrm{PS}}{\vdash} \alpha \Leftrightarrow W \vDash \alpha$$

For example, PS can be natural deduction and W can be interpretation (assignment of truth values to propositional variables).

- Given its goals and what agent believes to be true about the world the agent then **acts**, resulting in changes to world to obtain a new world W' and then senses new facts Δ_1 about world W' so to update its knowledge base:

$$\mathrm{KB}_{\Delta_0 \cup \Delta_1} \underset{\mathrm{PS}}{\vdash} \alpha \Leftrightarrow W' \vDash a$$

- Given what the agent now believes (all α it can infer) and goals, it can now act and sense again.

Can logic provide a complete account of all cognition?

Two reductionist arguments:

1. Complete proof system for just first-order (predicate) logic can simulate all of Turing-level computation (all computable functions $\rightarrow$ achievement of any complex goal).
1. If intelligence can be formalised mathematically, and since mathematics can be given logical foundations, then intelligence can be formalised in logic.

4.1.2. Monotonic and Non-monotonic Logic

The mathematics model/world is fixed/unchanging and axioms describing the world are specified without worry that they can later be contradicted when adding more facts (axioms). Hence classical logic is **monotonic**.

$$\mathrm{KB}_{\Delta_0} \underset{\mathrm{PS}}{\vdash} \alpha \rightarrow \mathrm{KB}_{\Delta_0 \cup \Delta_1} \underset{\mathrm{PS}}{\vdash} \alpha$$

However, real world is way too large and complex to encode all possible exceptions in our beliefs, so we use rules of thumbs and withdraw beliefs when we discover exceptions.

$$\mathrm{KB}_{\Delta_0} \underset{\mathrm{PS}}{\vdash} f \text{ but } \mathrm{KB}_{\Delta_0 \cup \Delta_1} \not\underset{\mathrm{PS}}{\vdash} f$$

e.g. where Δ_0 contains rule that all birds typically fly, Tweety is a bird so can fly, f, and Δ_1 contains newly sensed fact that Tweety is a penguin so we withdraw the conclusion and now infer $\neg f$.

Hence, the issue with classical logic is that it is *monotonic*, everything inferred from a set of axioms continues to be inferred after adding to said axioms. The only (impractical) solution is to write all the exceptions in the rule, which can be unwieldy.

This has led to development of non-monotonic logic for AI, essentially supplementing classic logic with defeasible rules of the form:

- birds *typically/usually (rather than necessarily)* fly
- birds fly *unless there is evidence to the contrary*

4.1.3. Model Theory / Semantics for Logic

The **model theory** (semantics) for *first order/predicate classical logic* typically consists of an interpretation:

- Domain of individuals D picked by/represented by constants in the language (e.g. tweety)
- n-ary tuples of individuals for which each n-ary predicate symbol P is true (e.g. interpretation of unary predicate fly is $\{(\text{tweety})\}$)
- for each n-ary function, mappings from sets of n individuals to single individuals (e.g. unary function symbol $\text{succ}(\text{deez}) = \text{nuts}$)

The **model theory** (semantics) for *non-monotonic logic* consists of preferred interpretations (models) $\rightarrow$ preferential model semantics.

Given two models in which interpretation of unary predicates *bird* and *penguin* is $\{(\text{tweety})\}$, the one in which interpretation of *fly* is $\{\}$ is preferred to the one in which it is $\{(\text{tweety})\}$.

4.1.4. Modal Logic

Modal logic extends classical propositional and predicate (first order) logic with operators expressing **modalities**. A **modal**, a word expressing modality, qualifies a statement.
e.g. "Sanjay is *usually* happy"

- Alethic modalities in Modal Logic (modalities of truth - possibility and necessity)
  p = it will rain today
  $\Box p$ = it's necessarily the case that in all possible worlds it will rain today
  $\Diamond p$ = possible in at least one possible world it will rain today

  Example of an interaction axiom: $\Diamond p = \neg\Box\neg p$
- Deontic modalities in Deontic Logic
  Oq = q is obligatory
  Pq = q is permitted
  Fq = q is forbidden

4.1.5. Beliefs, Desires and Intentions (BDI)

BDI logic describe the mental attitudes of agents acting in the world:

- **Beliefs**: about the world, the agent itself, and other agents.
  May not necessarily be true.
- **Desires**: motivational state of the agent, what we would like to accomplish.
- **Goals**: desires actively pursued by the agent (must be consistent unlike desires).
- **Intentions**: high-level plans put in place to achieve a chosen goal

We make intuitively reasonable inferences from axioms of BDI logics, e.g. $\text{Intend}\ q \rightarrow \text{Bel}\ \neg q$

4.2. Machine Learning Approaches to AI

Logic-based AI did not give particularly impressive results. In recent years, machine learning delievered much better results using algorithms such as back-propagation, multi-layer networks, availability of massive training data sets, massive increases in computing power, and specialised hardware.

>A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.

Overview of machine learning approaches to AI:

- **Supervised Learning**: builds mathematical model from data containing both input and desired output.
- **Semi-supervised learning**: models built from incomplete training data. Portion of sample input doesn't have any labels.
- **Unsupervised learning**: models built from data which contains only inputs and no desired output labels. Generally used to find structure in data, like grouping or clustering.
- **Reinforcement learning**: algorithms given positive/negative feedback in dynamic environment. Mimics training animals through punishment/reward.

Most successful models are artificial neural networks. Deep learning using multi-layer neural networks are responsible for many recent successes in AI.

4.3. Limitations of Logic and ML

Limitations of Logic-based GOFAI

- Too many rules need to be encoded to be useful, doesn't scale!
- Small amount of damage leads to complete failure — 'brittleness'.
- Inherently not designed to do low-level sensory tasks - problem of learning/acquiring data in practical way is not addressed.
- Symbol grounding problem — constants, functions, predicates — typically hand-crafted rather than grounded in data from real world. Semantics rely on meanings provided by designers.

Limitations of Machine Learning

1. problems with generalisation/transfer — trained network performs well on one task but poorly on new tasks, even if it is similar to originally trained data
2. problems with high-level congitive processes such as planning, casual reasoning, analogical reasoning (that one case or situation is similar to another).
3. **black box problem**: reasoning is opaque even to designers of the system
4. lack of symbolic/language like explanations of reasoning mean that one also cannot integrate reasoning of humans and machines so that they can jointly reason together

In more detail:

- Need very large data sets while humans learn from small amounts of data.
- Litle progress on new representations at levels of abstraction higher than input vocabulary.
  e.g. in vision, complex concepts unnecessarily difficult if we work from pixels and input representation, we need the intermediate concepts such as different objects!
- Deep learning has no natural way to deal with hierarchical structure — sentences are just sequences of words whereas human language is hierarchical.
- Deep learning struggles with open-ended inference.
  e.g. cannot draw inference about who isleaving who, or what will happen next with subtle difference between: "John promised Mary to leave" and "John promised to leave Mary"
- Deep learning thus far not well integrated with prior knowledge.
- Deep learning cannot inherently distinguish causation from correlation.

Problem of common-sense reasoning

- Fundamental problem for both ML and logic-based AI is common-sense reasoning
- e.g. we see a bird and don't explicitly check all possible exceptions before concluding if it can fly
- To some extent, we can solve this with non-monotonic logic but:
  1. often complex and not easily understandable to humans
  2. computationally not feasible if we want to ensure rational outcomes
     (need to check inference is consistent with everything else that is believed)
  3. developed for reasoning by individuals not by multiple agents reasoning together

4.4. Argumentation and Communication

Argumentation is a way of doing non-monotonic reasoning that is more in line with how humans reason and hence more understandable. It can be formalised to give rational outcomes even when computational/cognitive resources are limited.

It also allows integration of reasoning of multiple agents which will be later relative for ethical AI.

Essentially it is rationally choosing amongst (key aspect of intelligence):

- conflicting beliefs, or
- conflicting desires to adopt as goals, or
- alternative actions for achieving a goal

4.4.1. Non-monotonic reasoning as argumentation

Example

The proof of $\neg f$ from p and $p \to \neg f$ can be understood as an argument consisting of the argument's grounds/premises p and $p \to \neg f$ that support/justify the claim $\neg f$.

$$\text{A1} : (\{p, p \to \neg f\}, \neg f)$$

If we also consider the following then we have that each argument is a counter-argument to (attacks) the other since they have contradictory claims.

$$\text{A2} : (\{b, b \to f\}, f)$$

If we prefer A1 to A2 because of specificity principle (properties of subclasses take priority over super-classes), then A2 cannot attack so argument A1 wins out (claim $\neg f$).

We can identify inferences obtained by proof theory of non-monotonic logic:

$$\begin{array}{ll} b & \\ b \to f & \nvdash f \\ p & \vdash f \\ p \to \neg f & \end{array}$$

By constructing the arguments from the set on the left.
And possibly, using preferences, determine the winning arguments.

Example using logic programming

We can do this for many non-monotic logic including logic programming, e.g. Prolog.

Suppose:
1. **Believe** birthday $\wedge$ **Desire** celebrate_bday $\wedge$ **Believe** $\neg$ closed_cinema $\to$ **Goal**(cinema)
2. **Believe** birthday $\wedge$ **Desire** celebrate_bday $\wedge$ **Believe** $\neg$ closed_restaurant $\to$ **Goal**(restaurant)
3. **Believe** birthday
4. **Desire** celebrate_bday

Neither argument wins out (genuine dilema): **A1** (3,4,1) cinema $\longleftrightarrow$ **A2** (3,4,2) restaurant

Now suppose a rule:
5. **Goal**(restaurant) $>$ **Goal**(cinema)

A2 is stronger (preferred to) A1 so A1 cannot be moved as a counter-argument to A2. Hence an attack from A1 to A2 is cancelled out, and only A2 attacks A1 so A2 wins!

A1 (3,4,1) cinema $\longleftarrow$ **A2** (3,4,2) restaurant

Now suppose the rule:
6. closed_restaurant

A2 is now attacked by A3 as A2 assumes that closed_restaurant is not provable. Hence A1 wins.

A1 (3,4,1) cinema $\longleftarrow$ **A2** (3,4,2) restaurant $\longleftarrow$ **A3** (6) closed_restaurant

Visual examples

Example 1

$$\begin{array}{lll} 1.p & & \\ 2.p \land \neg q \to r & \vdash r & \textbf{A1}\ (1,2)\ \checkmark \end{array}$$

Example 2

$$\begin{array}{lll} 1.p & & \textbf{A1}\ (1,2)\ \times \\ 2.p \land \neg q \to r & \nvdash r & \uparrow \\ 3.s & \vdash q & \\ 4.s \land \neg g \to q & & \textbf{A2}\ (3,4)\ \checkmark \end{array}$$

Example 3

$$\begin{array}{lll} & & \textbf{A1}\ (1,2)\ \checkmark \\ 1.p & & \uparrow \\ 2.p \land \neg q \to r & \vdash r & \\ 3.s & \nvdash q & \textbf{A2}\ (3,4)\ \times \\ 4.s \land \neg g \to q & \vdash g & \uparrow \\ 5.f & & \\ 6.f \to g & & \textbf{A3}\ (5,6)\ \checkmark \end{array}$$

4.5. Dung's theory of Argumentation

Given an **Argument Framework** $\langle \text{Args}, \text{Attack} \rangle$ where $\text{Attacks} \subseteq \text{Args} \times \text{Args}$, label each argument in Args with either IN (winning), OUT (losing), or UNDEC (undecided).

A labelling of Args is legal (valid) iff:

1. If $X \in \text{Args}$ is labelled IN then for every Y such that $(Y, X) \in \text{Attacks}$ Y is labelled OUT
2. If $X \in \text{Args}$ is labelled OUT then for every Y such that $(Y, X) \in \text{Attacks}$ Y is labelled IN
3. If $X \in \text{Args}$ is labelled UNDEC then for every Y such that $(Y, X) \in \text{Attacks}$ Y is labelled UNDEC and there is no such Y such that $(Y, X) \in \text{Attacks}$ and Y is labelled in IN

Let $L_1, ..., L_n$ be the legal labellings of $\langle \text{Args}, \text{Attacks} \rangle$.

1. There will always be a unique L_i that has a minimum number of arguments labelled in IN:

$$\exists L_i \forall Lj : \text{IN}(L_i) \subseteq \text{IN}(Lj)$$

   L_i is said to be the **grounded** labelling and the arguments labelled IN in L_i ($\text{IN}(L_i)$) are the grounded extension of $\langle \text{Args}, \text{Attacks} \rangle$.
2. There may be more than one labelling with a maximal number of arguments labelled IN:

$$\neg\exists Lj : \text{IN}(L_i) \subset \text{IN}(Lj)$$

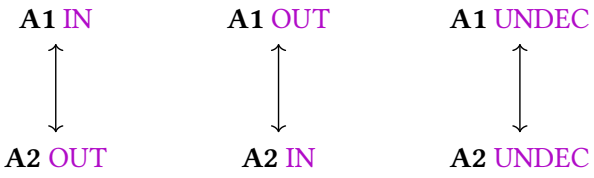
   Each such L_i is said to be the **preferred** labelling and the arguments labelled IN in L_i ($\text{IN}(L_i)$) are the preferred extension of $\langle \text{Args}, \text{Attacks} \rangle$.

Examples of labellings

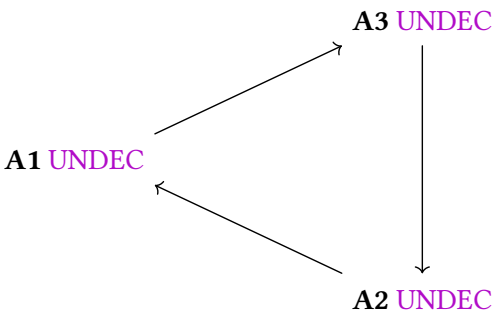
*One legal **grounded** and **preferred** labelling*

A1 IN ⟵ **A2** OUT ⟵ **A3** IN

Three legal labellings

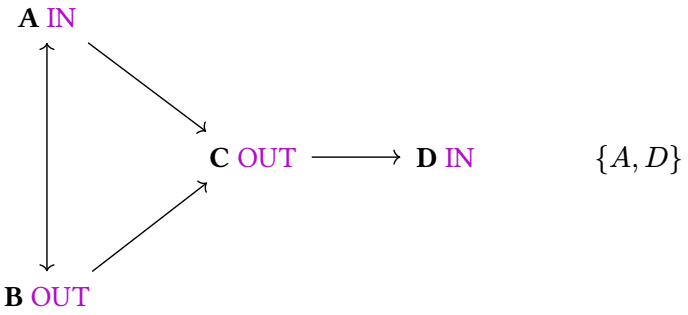


*One legal **grounded** and **preferred** labelling*

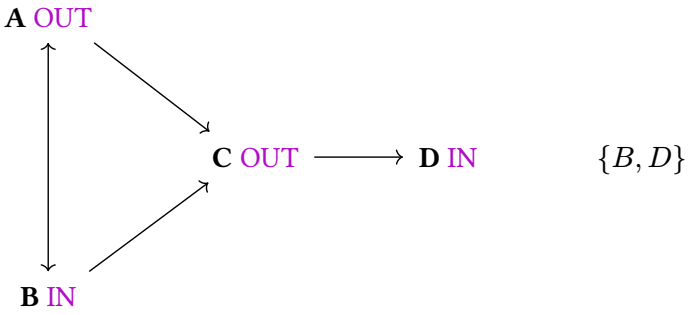


Further examples of labellings

*2 **preferred** extensions*

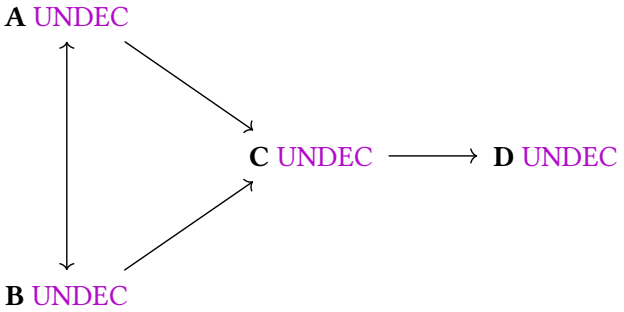


$\{A, D\}$



$\{B, D\}$

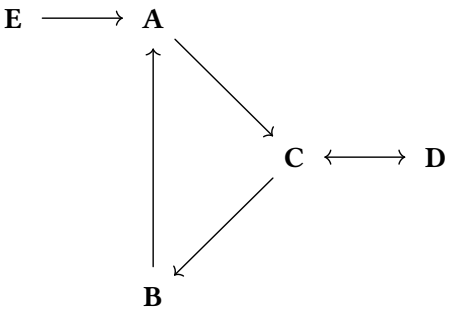
*1 **grounded** extension*



grounded extension $= \{\}$

*2 **preferred** and **1** grounded*

Preferred: $\{E, C\}$, $\{E, D, B\}$
Grounded: $\{E\}$



4.6. Argument Game Proof Theories

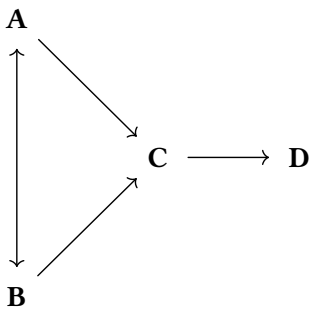
Given an argumentation framework $\langle \text{Args}, \text{Attacks} \rangle$ constructed from an agent's knowledge base: there are argument games defined for deciding whether some $x \in \text{Args}$ is in an s (= grounded or preferred) extension.

The agent plays the roles of PRO and OPP:

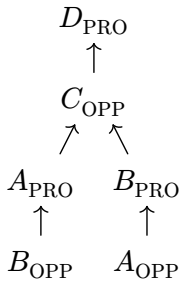
- PRO moves $\mathcal{X}$
- OPP and PRO take turns, referencing $\langle \text{Args}, \text{Attacks} \rangle$ to move attacking arguments against other player, subject to rules of the game that vary according to the criteria/semantics s.
- If PRO successfully counters each OPP move and OPP moves exhaustively (subject to game's rules) then PRO establishes members of $\mathcal{X}$ in an s extension.

Example

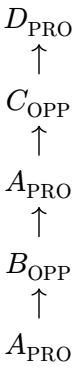
Suppose the arguments from earlier:



In a grounded game, PRO cannot repeat in any given path from root to leaf (dispute) but OPP can repeat in a dispute. PRO will lose this game (D is not in the grounded extension).



In a preferred game however, OPP cannot repeat in a dispute but PRO can repeat in a dispute. So PRO wins this game: D is in a preferred extension.



4.6.1. Distributing reasoning via dialogue

So far we have assumed a single agent constructing $\langle \text{Args}, \text{Attacks} \rangle$ from their private KB Δ and using s argument game to determine whether an argument is in an s extension of $\langle \text{Args}, \text{Attacks} \rangle$.

However, reasoning, especially about contentious beliefs/goals/actions, is often done with others. We need to jointly reason about complex issues integrating individual knowledge and expertise from many sources. We adapt argument games to obtain dialogical models of distributed/joint reasoning:

Formally, a s **dialogue game with public semantics** is adjudged to be in favour of an argument X (and its claim) $\Leftrightarrow$ X is in an s extension of the framework built from the contents of what have been publically communicated.

Example dialogue

Suppose two agents Ag1 and Ag2 exchanging messages:

Ag1	Ag2	$(\text{Args}, \text{Att})_\Delta$
		$q \land \neg s \to p$
$X = [q \text{ and } q \land \neg s \to p]$		m
	$Y = [m \text{ and } m \land \neg g \to s]$	$m \land \neg g \to s$
		m
$Z = [\neg t \to g]$		$\neg t \to g$

We incrementally define the knowledge base by contents of exchanged arguments.
Ag1 wins grounded dialogue game for argument X claiming p is true iff argument X claiming p winning under grounded criterion.

Essenitally, a dialogue that reasons that α is the case effectively demonstrates that α can be inferred from the contents of the arguments moved during the dialogue.

5. Ethics and Morality

5.1. Ethical Challenges for Artificial Intelligence

In the past:

1. AI machines were deployed in **segregated spaces** restricted to **trained personnel only**.
2. Used to be in **specific/predictable domains** with **limited autonomy** (ability to independently make decisions).
   Hence allowing for possible ethical/legal/societal issues to be handled during development.

These days, we find:

1. AI systems are packaged in off-the-shelf software and hardware available for customisation and deployment in unpredictable contexts — often 'non-segregated environments'
1. Greater autonomy is provided to AI systems as they are developed further — autonomous weapons / cars
2. There is a lack of transparency & explanation as to how ML algorithms make decisions

Examples of areas of ethical concern

- Autonomous cars; brake failures leading to decisions between injury of pedestrians or the driver
- R&D in autonomous weapons in U.S., UK, China, Russia, France, Israel, South Korea
- ML algorithms used in deciding insurance premiums, recruitment, mortgages, criminal justice; issues of bias, fairness, and transparency crop up
- sex robots & robots developed for care of elderly/disabled
- social media & IoT monitoring and analysing human activity; data privacy concerns
- filtering algorithms polarise beliefs
- targeted advertising/message manipulation
- medical decision support systems, surgical robots, robots assisting autistic children, image diagnosis
- superintelligent AI & the value loading problem

Ethics concerns what **we** *ought to do*. What 'the good' is, and how to ensure that our behaviour ensures good outcomes and avoids bad outcomes.
Typically, terms ethics and morality are used interchangeably.

In the context of AI, uses of, interactions with, and the decisions made by AI machines, have and will increasingly have an ethical impact on us. As AIs become more intelligent and autonomous, we need to think about ensuring AIs are designed to behave ethically.

Questions & issues raised when thinking about ethics of AI

1. We are the users and designers of AI, so we must think about:
   - our ethical uses of AI
   - designing AIs to ensure they cannot act other than ethically
   - how to make AIs choose ethically good courses of actions; especially as more autonomous AIs may have the capacity ot act unethically
2. AI systems are being increasingly embedded within technology so it is hard to distinguish which ethical issues are presented by the AI itself or other technology
3. AI systems are increasingly influencing our behaviours & social interactions
   How do we ensure we remain aware of their ethical impact?

5.2. Philosophical Accounts of Ethics and Morality

Below are three key theories of ethics/morality.

Deontology — focuses on duties and rights that are fulfilled and respected by obeying rules, norms, and laws; morality is about finding a good system of rules to guide our behaviour and sticking to these rules; it is concerned with choices that are **morally forbidden** (prohibited), **required** (obligated), or **permitted**

Consequentialism — focuses on actions that have the 'best' outcomes/consequences; we should asses the most likely results of our actions and choose the actions with the best results

Virtue Ethics — focuses on cultivation of good personatlity traits/virtues; persons with these virtues act in morally good ways; becoming more virtuous means becoming more ethical; we should work on becoming more honest, compassionate, kind, courageous, etc

Example of application of these key theories

Suppose it is obvious someone needs help:

- A *deontologist* would say that doing so (helping) would be in accordance with a moral rule such as "treat others as you would want them to treat you".
- A *consequentalist* would say that the consequences of doing so will maximise well-being.
- A *virtue ethicist* would say that helping the person would be the act of a charitable or benevolent person.

5.2.1. Deontology

Deontology is concerned with choices that are morally forbidden (prohibited), required (obligated), or permitted.

Deontological theories therefore guide and assess our choices of what we ought to do.
There are two broad approaches:

- those that focus on the *duties* of the agents that act to be moral
- those that focus on *rights*

In either case, we focus on following a set of rules that encode moral duties and a respect for rights.
For example, *you should not kill* encodes both a duty and a respect for the right to live.

> **Note**
>
> In contrast to consequentialism, some choices cannot be justified by their effects: no matter how morally good (bad) their consequences, some actions are morally prohibited (obliged).

Examples of Deontological rules

In religious moral codes:

- "you must not eat beef" (prohibition in Hinduism)
- Ten Commandments in Judeo-Christian religions, e.g. you must not commit murder, adultery

Secular (non-religious) moral codes:

- Legal encodings of morality — laws prohibiting murder, theft, tax evasion
- Asimov's laws of robotics

Some rules are more abstract and provide general principles that need to be 'appropriately' applied in any given situation:

- The Golden Rule — "Treat others as you would like others to treat you"
- Emmanuel Kant's (1785) Categorical (unconditional) Imperative (command) — "Act only according to that rule by which you can at the same time will that it should become a universal law."

  The **categorical imperative** (CI) commands that *every rule you act on* must be such that you are *willing to* make it the case that *everyone always acts according to the same rule* when in a similar situation.

  > **Note**
  >
  > Suppose you are considering whether to act according to the rule "if I want something, then it is permitted to lie to get that something". One would have to be willing that everyone lied to get what they wanted — but then nobody would believe you or rather nobody would believe anyone.
  >
  > - So, if you **willed** lying to become a universal law, you would not achieve your goal.
  > - Therefore according to the CI, it is **not** permitted to lie.
  > - It is not permitted because the only way to lie is to make an **exception for yourself**.
  >
  > Kant believed this is a *purely rational recipe* for how to act — if you don't act according to it, then your goals will be blocked. An equivalency is provided to the more explicit moral: *"Act so that you treat humanity (yourself and others) always at the same time as an end and never simply as a means."*

5.2.1.1. Critiques of Deontology

Common critiques levied against deontology:

1. Too general/vague and so doesn't always guide action in a specific situation
2. Inflexible — provides us with absolute prohibitions on certain actions
   What about lying to save someone's life? Is it permissible to kill in self defense?
   These rules could be updated in principle to include exceptions to create new rules that we would want everyone to follow.
3. Deontologists usually fail to specify which principles should take priority when rights & duties conflict.
   Deontology thus cannot provide complete moral guidance
4. Applications of rules mean that overall **consequences** of actions may be harmful
   Consider a sadist as a masochist who follows the Golden Rule. Consider that CI only allows killing as a side effect of doing something for the greater good and not as a means to achieve greater good — you cannot kill one person to save any number of people.

5.2.2. Virtue Ethics

Virtue Ethics describes the character of a moral agent as a 'driving force' of ethical behaviour.
(its origins lie in ethical theories of Socrates and Aristotle)

Virtue Ethics says that morally good actions flow from the cultivation of good character, which consists in realisation of specific virtues.
e.g. courage, truthfulness, modesty, generosity, wisdom, justice

What is 'good' is learnt from particular actions, from making connections between actions and their consequences. Humans learn what is good through intuition, induction, and experience. (e.g. by asking good people about the good, one's sense of the overall good comes into focus, and the ideal person acquires practical wisdom and moral excellence)

Critiques of Virtue Ethics:

- Different people, cultures, and societies often have different opinions on the constitution of a virtue, and it can be argued there is no objectively right list.
- Does not provide guidance on what sorts of actions are morally permitted/diallowed, but rather qualities someone should cultivate to become a good person.

5.2.3. Consequentialism

Consequentialists say that actions should be morally assessed only by the states of affairs they bring about, specifying which states are good, and then saying that whatever actions increase Good are morally right — consequences of one's actions are the ultimate basis for any judgement about rightness/wrongness of those actions.

Consequentalists often differ widely in specifying Good, though one of the most popular interpretations is defined in terms of 'happiness'.

5.2.4. Moral Relativism

The **moral relativism** argument is that: Given (1) morality is simply the expression of ethical or value judgements we make in a given society and (2) other societies have their own judgements, we have ours. Therefore we should not judge other cultures.

However, the conclusion itself is a 'value judgement', that is to say making a judgement about what one should/n't do, as if it were a universal truth. This is a contradiction.

5.3. Implementing Moral Agents

Moor's Categorisation of Moral Agents

- *Ethical Impact Agents*: any machine that can be evaluated for its ethical impact if the agent's operations increase or decrease good in the world
- *Implicit Ethical Agents*: machines *designed* to not have negative ethical effects — behaviours constrained by designers following ethical principles
- *Explicit Ethical Agents*: machines that can reason what the best action is in ethical dilemas and novel situations not anticipated by using ethical principles
- *Full Ethical Agents*: machines that can be said to have full 'moral' agency and are able to justify their moral judgements

There are three approaches to implementing Artificial Moral Agents (AMMs):

- *Top Down*: existing moral theory (ranging abstract to specific) is chosen and encoded in the agent which applies the theory in concrete situations to act ethically — typically logic-based and deontic
- *Bottom Up*: agent explores course of action, learns, and is rewarded for behaviour that is morally praise-worthy (reinforcement learning of ethical values)
- *Hybrid*: use elements of both

5.3.1. Top Down Implementation of Deontic AMAs

Deontic logic is well-studied and support by a wide variety of programming frameworks.
Recall deontic modalities in deontic logic:

- $\mathrm{Oq} = q$ is obligatory
- $\mathrm{Pq} = q$ is permitted
- $\mathrm{Fq} = q$ is forbidden/prohibited

Example: Build a Deontic agent for an autonomous car

Let's start with these two rules:

- F drive on right
- O protect driver

Let's now suppose we are in a two-way tunnel and a collision is anticipated as highly likely, with a caculated risk of injury to the driver as high. We must swerve and drive on right, we need to prioritise an obligation:

1. collision immiment $\wedge$ risk estimation (injury, left, X) $> 0 \rightarrow$ O protect driver
2. F drive on right

... but only if the risk of injury due to collision with oncoming traffic on risk is less than risk of injury to stay on the left:

1. collision immiment $\wedge$ risk (injury, left, X) $\wedge$ risk (injury, left, Y) $\wedge\, Y < X \rightarrow$ O protect driver
2. F drive on right

... but what if the car is not in a tunnel, etc, list goes on.
Only way to guarantee the correct conclusion is to explicitly list all possible circumstances and exceptions to apply and prioritise.

So rule-based/deontic logic specification of correct behaviour is okay if agent is deployed in a very restricted environment and the rules encode all forseeable situations. But for autonomous agenjts in complex and unpredictable environments, deontic specifications of ethical behaviour need to be more general. *This presents challenges such as common-sense interpreting intentions underlying principles, prioritising conflicts amongst rules, avoiding deadlock, avoiding unforseen consequences.*

Example: Asimov's Four Laws of Robotics

0. A robot may not injure humanity, or, by inaction, allow humanity to come to harm (a generalisation of law 1)
1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws

We may ask:

- What does it mean to harm a human?
- Literal minded robot following Law 0 might interrupt a surgeon about to cut a patient.
- Is context fully described and understood? — what counts as harm?

5.3.2. Implementation Details

How would an AMA implement Kant's Categorical Imperative?

A powerful AMA, when considering an action according to rules that are either programmed from the outset or learnt over time, runs a simulation to determine whether its goal would be blocked if all other agents were to operate according to the same rule. (requires great understanding about conext, casual reasoning, human psychology, ...)

Bottom up machine learning approach to implementing AMAs can be equated with virtue ethics, both emphasise development of ethical behaviours through training and learning, though perceiving how actions lead to good ends without explicit symbolic rules governing behaviour.

5.4. Utilitarianism

Utilitarianism is a specific definition of consequentialism that defines happiness as the goal, i.e. a moral philosophy that impartially promotes actions that maximise happiness.

Why use happiness?

- True for everyone — basis for universal morality.
- Hard to think of values worth valuing (depends on society/culture, not ultimately linked to happiness). Can be related to personal relationships, virtues, pursuits, and governance.

What is happiness?

1. Universal underlying value
2. Subjective experience of well-being
3. If we think 'counter-factually', the absense of something that reduces our happiness is the thing that makes us happy.

5.4.1. Trolley-ology (as an argument against Utilitarianism)

Slides 35-43 provided in `7ccsmeai-trolley-ology.pdf`.

Conclusions: a combination of physical force and means to an end that is important and interferes with utilitarian calculation. Why?

1. We have evolved to have emotional reluctance to violence as it disrupts cooperation in society.
2. Instinctive distinction between means to ends and side effects interferes with our rational utilitarian calculations. Likely because when we think about plans of action our emotional alarm system responds to harms that we anticipate, and means that are casually necessary for achieving goals are more represented cognitively, unlike side-effects of our actions.

   In international law, it is illegal to intentionally bomb civilians to lower enemies moral (used as a means) but legal to bomb munitions factory knowing full well that civilians will die (side effect).

Our emotions/moral intuitions are generally sensible but not always correct ← we allow emotions/moral intuitions to influence our ethical decisions and arguably make wrong decisions.

Some use the trolley-ology thought experiments to show there is something wrong with utilitarianism, as in some situations people apply utilitarian calculations but in others they don't. Emotional decision isn't necessarily the right decision.

Future AIs will not have utilitarian reasoning biased by evolution.

5.4.2. Other supposed arguments against Utilitarianism

1. Utilitarianism is not demanding enough and sometimes leads to intuitively unethical choices.
   Suppose five people all die because they don't receive organ transplants. Utilitarianism says it would be ok to harvest organs from another person who would then die to save the five. What kind of society would be better/happier? One in which this is allowed/or not allowed?
2. Utilitarianism is too demanding and requires sacrifices that are unreasonable.
   Suppose you had to ruin a 500$ suit to save someone. Utilitarianism says you should've donated that money instead of spending it on a suit.
   Pragmatic utilitarianism says being a perfect utilitarian is anti-utilitarian since it is incompatible with the way our brains are designed.

6. Algorithms

6.1. Accountability and Transparency

A lot of issues arise with algorithmic decisions:

- How can we ensure accountability when algorithms make decisions?
  how and to whom/what do we assign Responsibility/blame when things go wrong
- Many important decisions now made by computers.
  Algorithms count votes, approve loan and credit card applications, target citizens, or neighbourhoods for police scrutiny, select taxpayers for auditing, grant or deny visas, …
- Tools available to policymakers, legislators, and courts developed to oversee human decision making but these often fail when applied to computers.
  How do you judge intent of software?
- Automated decision systems can return potentially incorrect, unjust, or unfair results.
  Additional approaches needed to make systems accountable and governable.

So how can we ensure accountability? To demonstrate to public at large and to regulatory/legal bodies that automated decisions comply with key standards of legal/ethical fairness?

Transparency is the ability to examine and explain how algorithms make decisions which is important in ensuring accountability. It also allows one to demonstrate ethical fairness and provide incentives to ensure compliance with ethical standards.

Application to AI

In Plato's Republic, Protagoras considered a ring ('Ring of Gygnes') that makes the wearer invisible, arguing that the wearer would commit 'all manners of wrong-doing'.

There are fears that without our knowledge, we may be manipulated by powerful machines or corporations taking advantage of opaque AI systems.

Transparency, hence accountability, is said to reduce the likelihood of this.

Sometimes transparency is not desirable:

- Explaining a decision may reveal private data about an individual (violating GDPR).
- Transparency may be abused to game systems such as tax auditing or airport screening.

Demanding transparency involves weighing up the harms and benefits - transparency isn't a good fit to be a universal value/virtue or deontological rule.

When AI algorithms are used in narrow and predictable contexts, a basic requirement is **procedural regularity**: each participant will know that the same procedure was applied to them and that the procedure was not designed in a way that disadvantages them specifically

Tools for procedural regularity ensure that:

- Same policy/rule was used to render each decision
- Decision policy was fully specified before decision subjects were known
- Each decision is reproducible from specified decision policy and inputs
- If a decision requires any randomly chosen inputs, they are beyond control of any interested party

Techniques for ensuring procedural regularity:

- *Software verification* involves logic-based techniques to formally prove the program's behaviour satisifes certain properties (invariants) or facts about a program's behaviour that are always true regardless of internal state or input data.
  We could model check the program - exhaustively check all inptus to ensure invariant is not violated.
- *Cryptographic commitments* can be used to ensure the same decision policy was used for each of many decisions.

ACM Principles for Algorithmic Transparency and Accountability

- *Awareness*: owners, designers, builders, users, and other stakeholders of analytic systems should be aware of possible biases involved in their design, implementatrion, and use, and potential harm that biases can cause to individuals & society
- *Access and Redress*: regulars should encourage adoption of mechanisms that enable questioning and redress for individuals and groups that are adversely affected by algorithmically informed decisions
- *Accountability*: institutions should be held responsible for decisions made by algorithms that they use, even if it is nto feasible to explain in detail how the algorithms produce their results
- *Explanation*: systems and institutions that use algorithmic decision-making are encouraged to produce explanations regarding both the procedures followed by the algorithm and specific decisions that are made
- *Data Provenance*: document inputs, entities, systems, and processes that influence dat aof interest (e.g. training data) and so provides a historical record of the data's origin, what processing it underwent (e.g. providing links to algorithms it was processed by, together with input parameters), and the systems/entities that processed the data.
  Such a record should be maintained and ideally accompanied by an exploration of potential biases introduced by human/algorithmic data-gathering process.
  Public scrutiny of data provides maximum opportunity for corrections.
  This does however raise concerns over privacy, protecting trade secrets, or revelation of analytics that might allow malicious actors to abuse the system. This can justify restricting access to the system to qualified and authorised individuals.
- *Auditability*: models, algorithms, data, and decisions should be recorded so that they can be audited/inspoected in cases where harm is suspected
- *Validation and Testing*: institutions should use rigorous methods to validate their models, and document those methods and results (routinely perform tests to assess and determine whether the model generates discriminatory harm)

6.2. Liability and Responsibility

Accountability means institutions should be held responsible for decisions made by the algorithms that they use, even if it is not feasible to explain in detail how the algorithms produce their results.

Application to autonomous vehicles

With increasing autonomous AI systems, who or what is responsible (liable) for when things go wrong? Driver error causes 94% of conventional car crashes. Fully & partially autonomous vehicles could improve that number, however we will still see accidents involving AVs, so who is responsible?

The current assumption is that the companies behind the software/hardware are legally liable and not the car owner or the person's insurance company. To date, several accidents show an AV warning the driver to disengage autopilot and take manual control, does this then shift the responsibility to the human? Would they even be given adequeate time to react? Some cars may not even have a driver or a steering wheel?

There is one (and only) recorded case of a fatal AV accident in Arizona (March 2018) where the initial police inestigation indicated the pedestrian at fault.

Read more here: https://en.wikipedia.org/wiki/Death_of_Elaine_Herzberg

Could an AI system be held **criminally liable**? Typically this requires attributing an action and mental intent, a conscious understanding of doing the action and that it leads to harm.

1. **perpetrator via another** applies when an offense has been committed by a mentally deficient person or animal, who is therefore considered innocent. Anybody who instructed the offender can be held criminally liable, e.g. a dog owner instructing the animal to attack.

   An AI program could be held to be an innocent agent, with either the programmer or user being held to be the perpetrator-via-another.

2. **natural probable consequence** means when one can anticipate a result that is reasonably expected to occur from an action. If someone does something that is likely to cause harm or break the law ("negligent") then they can be assumed to have intended that harm or illegal outcome.

   In 1981, an AI robot in a Japanese motorcycle factory killed a human worker. The robot wrongly identified the employee as a 'threat to its mission', and calculated the most efficient way to eliminate 'the threat' was by pushing him into an adjacent operating machine. We want to understand whether the programmer knew this outcome was a probable consequence of its use.

3. **direct liability** applies when there is both action and intent. Action is straightforward to prove, but intent can pose a problem.

   There are also cases where criminal liability may not apply, the matter would have to be settled with civil law, at which point we have to ask whether an AI system is a service or a product:
   - *product*: product design legislation applies; based on warranty
   - *service*: rules of negligence apply; plaintiff must demonstrate three elements to prove negligence (defendant had a (deontological) duty of care, defendeant breached that duty, and breach caused an injury to the plaintiff)

6.2.1. Moral Responsibility, Free Will, and Justice

With respect to philosophy and law, understanding whether an agent can be said to be *morally responsible* is crucial in deciding whether the agent deserves praise, blame, reward, or punishment for its action.

Typically, a person is morally responsible for a crime if:
1. they were conscious ('understood') that the crime would cause harm (understood ethical impact)
2. was free to do otherwise — exercised their *free will* in choosing to commit the crime rather than not commit the crime

   Assumption of free will is tied with idea that there is a 'self' that chooses freely.

The **self**: "the I that I am - the you that you are"; the sense one has that they are more than just a physical body; that there is a self that you, thinks, is the subject of conscious experience (suffering, joy, pleasure); an identifiable self that exists and persists over time; that exercises free will.

Conditions for moral responsibility
- conscious understanding of ethical impact
- freedom to choose otherwise

These also characterise Moor's notion of a full ethical agent.

Moral responsibility means that the law can exact a penalty on a person (imprisonment, fines, etc) in order to:
1. *protect* society from the criminal, and protect the criminal from society
2. *deter* others from committing the crime
3. *rehabilitate* the criminal so they don't commit the crime again
4. punish the person (*retribution*)

Retribution is punishing the *self* for having *freely* chosen to commit a crime.

> **Note**
>
> Would there be a *distinct AI that freely chooses amongst options*?
> - In the current state of things, it is the case that AI is just algorithms that process information about the world, the effects of actions, and then makes choices that maximise utility / follow a set of rules / actions on basis of rewards.
> - Decision as to how one judges/evalutes utility of states is algorithmic decision
> - There is no *self* in the AI
>
> We may conclude that **there is no self, there is no free will, on the basis of which we can say that an AI has moral responsibility**.
>
> We could instead use a consequentalist framework to decide penalties in order to ensure:
> 1. *protection* of society from AI
> 2. *deter* others' AIs from committing same crime
> 3. '*rehablitate*' the AI so it doesn't commit the crime again
>
> Penalties need to ensure action selection modules of other AIs (deterrence) and criminal AI (rehab) are appropriately modified so no criminal action is chosen.
>
> We find that retributive punishment doesn't make sense as there is no *self* that *freely* chose the action, and hence would deserve punishment.

> **Note**
>
> What about if we apply this logic to humans? Does the same reasoning apply?
> Is there a *distinct human self that freely chooses amongst options*?
>
> It is common to think of the self as the 'chief executive' overseeing workings of the mind - the I (ego/self) is the subject of experiences, has thoughts, freely chooses, …
>
> *Some things people might think of themself:*
> - 'I was once very small'
> - 'If there are no major accidents, I will become old'
> - 'My body may change but I will remain the same'
> - 'I might have been born at another time or place'
> - 'It would be great to be Leonardo Dicaprio for a day'
> - 'My self will survive after my body has died, to ascend to heaven or born into another body'
>
> It is common to think of the self ('the soul') as separate from the body.
>
> Closely examining the idea of a self, we only find neurons firing, giving rise to thoughts and experiences. *Where is the distinct inner self that has these thoughts and experiences?*

6.2.2. The Self is an Illusion

Modern neuroscience, philosophy of the mind, and Buddhism all converge on the idea that the notion of a distinct self - someone looking out at a world out there, and also watching our own thoughts pass by - is **an illusion**.

There are only mental processes: streams of thoughts, sensations, and perceptions passing through our brains, but no central place where all of these phenomena are organised. No "CEO"/self that weaves streams together to create a unified integrated experience.

> **Note**
>
> It can be argued that when Descartes said
>
> I think therefore I am
>
> he should've actually said
>
> Thinking is being

The **sense of self** is how we consciously experience the integration of information.

For evolutionary reasons, it makes sense to have developed a distinct self as it is crucial to our identity / who we are, and drives self preservation which implies body preservation which in turn ensures genes are passed down.

6.2.3. Free Will is an Illusion

Free will is also **an illusion** if we consider that the laws of physics tell us everything is determined by what has gone before, it is all a chain of cause and effect. Given a choice of A and B, the choice made is determined by neuronal processes in the brain, which in turn obey laws of physics. At a higher level, the choice is fully determined by our genes and environment (sum of experiences).

> **Note**
>
> This doesn't mean we *don't have choices*, it's just that they are ultimately decided by what has happened before (environment history), our genes, and ultimately explainable in terms of the laws of physics in which every event that takes place in tyhe world is determined by preceding chain of cause and effect.
>
> The conscious experience of making a choice - what it feels like when we make a choice - is the feeling that we freely make the choice.
>
> So, free will as its commonly understood, as 'being really free' can only mean a choice that is not at some level determined by the law of physics otherwise it would not be commonly understood as free will.
>
> *Free will is a concept that is only meaningful when we refer to what it feels like when we make a choice.*

What does this mean for justice for humans?

Retributive justice is about exacting revenge against the self that freely made a choice, but if there is no self freely making a choice, then we should abandon the idea of retribution, and focus on other reasons for penalising criminals (protecting society from criminal and vice-versa, detering others, rehabilitating the criminal).

These reasons are effectively consequentalist (ultimately utilitarian; penalities are imposed to maximise well-being). Penalties act on the choice function of the criminal (rehabilitation) and others (deterence), to ensure right choices are made in the future.

If we treat people as if they had selves/free will (so are morally responsible) much like we think of ourselves as having selves/free will, then we can feel justified in penalising them.

Caveat: for more serious crimes, it may be that if a criminal is not punished, society could break down due to a lack of faith and trust in the criminal system, so there may be utilitarian arguments for having some retribution.

In practice, what does this mean for AI systems?

- They, like us, do not have distinct real selves / free will
- Issue of moral responsibility is irrelvant to deciding how we should react when things go wrong and we/they make wrong choices
- **We should adopt a consequentalist/utilitarian approach** - penalities are such that we ensure:
  - AI is taken out of use so it doesn't commit crime again
  - Penalties are imposed on companies, corporations, and institutions to deter development and use of AIs that commit similar crimes

6.3. Algorithmic Bias

Bias in algorithmic systems describes systematic and repeatable errors in a computer system that creates unfair outcomes, such as favoring one arbitrary group of users over others.

> Example
>
> For example, a credit score algorithm may deny a loan without being unfair if it is consistently weighing relevant financial criteria.
>
> If an algorithm recommends loans to one group of users, but denies loans to another set of nearly identical users based on irrelevant criteria, and this behaviour can be repeated, then an algorithm can be described as biased.

6.3.1. Types of Algorithmic Bias

- **Data Bias** — original training data is biased so the algorithm will perpetuate and potentially compound biases
  A key problem is that algorithms trained on data sets reflect societal biases at time of data collection. Society evolves and changes so what is considered 'biased/unfair' may change.
  - *Thinking about statisical standards:* there may also be environmental issues, e.g. training data from different geographical areas may not apply in other geographical areas.
  - *Thinking about moral standards:* biases may be relative to moral standards, e.g. an algorithm allocating money for health spending might allocate based on success but less successful treatment might be in more under-privileged sub-population, and it may be less successful because they are under-privileged, so one probably shouldn't prioritise spending on the more successful treatment.
- **Algorithm Bias** — bias in the design of an algorithm
  e.g. a search engine only showing three results per screen can be understood to privilege top three results more than next three
- **Use and Interpretation Bias** — use of an algorithm's output can be subject to bias:
  - use in unintended contexts; e.g. autonomous vehicle used in env. it was not trained for
  - interpreting outputs as objectively correct; e.g. probation officer who uses algorithm to predict risk of re-offending belives algorithm is objective/infalliable, so automatically accepts suggestions without considering alternative assessments from own knowledge and experience
- **Feedback Bias** — many algorithms are further trained and optimised over time, usually based on data concerning whether original recommendations or predictions were judged accurate or useful by human operators; hence any of the prior mentioned biases can be further amplified over time

6.3.2. Problems with identifying algorithmic bias

1. *complexity* — analysing algorithmic bias has grown in difficult alongside complexity of programs/ design; decisions made by one designer, or team, can be obscured among many pieces of code created for a single program; decisions may be forgotten over time; collective impact of design on program's output may be forgotten
2. *black box problem* — explaining how ML algorithms make decisions is an unsolved issue

6.3.3. Recent Examples of Bias

1. COMPAS algorithm widely used in U.S. to guide sentencing by predicting criminal reoffence was revealed to be racially biased in 2016. Actual re-offending rates did not match up with the predictions, biased against black defendants and biased for white defendants. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-
2. In the U.S., PredPol algorithm predicts when and where crimes will take place, aim to reduce human bias in policing. Found in 2016 to be unfairly targeting certain neighbourhoods.

   In simulation, it was found that there was a feedback loop that exhacerbated racial biases as the algorithm was trained from police reports rather than the true crime rates. So as officers were sent to neighbourhoods with high proportion of people from racial minorities, the software would indicate they should keep going there since they frequently go there. https://www.newscientist.com/article/mg23631464-300-biased-policing-is-made-worse-by-errors-in-pre-crime-algorithms/
3. In 2018, researchers at MIT found three of the latest gender-recognition AIs (IBM, Microsoft, Megvii) could correctly identify a person's gender from a photograph 99% of the time - but only for white men. Dark-skinned women accuracy dropped to 35%. Increases the risk of false identification of women and minorities. Likely due to a bias in the data containing more white men than black women. https://www.newscientist.com/article/2161028-face-recognition-software-is-perfect-if-youre-a-white-man/
4. In October 2017, police in Israel arrested a Palestinian worker who had posted a picture of himself on Facebook posing by a bulldozer with caption 'attack them' in Hebrew. Except in reality it was a mis-translation from 'good morning' as it is very similar in Hebrew & Facebook's automatic translation software chose the wrong one. Man was questioned for several hours before someone spotted the mistake. https://www.theguardian.com/technology/2017/oct/24/facebook-palestine-israel-translates-good-morning-attack-them-arrest

6.3.4. Solutions to problem of algorithmic bias

1. *statisical approaches*:
   - pre-processing / modification of training data with aim to prevent algorithm from learning discriminatory decision-making rules in training stage
   - in-processing involving modifying algorithmic model itself, e.g. changing crtieria that result in 'branches' in a decision tree in order to ignore/correct influence of protected characteristics (many in-processing methods require personal data regarding protected characteristics which may not be available)
   - post-processing to remove discriminatory rules or otherwise modifying the model after it has been trained
2. *discursive frameworks, self-assessment tools, and learning materials*: number of tools have been developed for self-assessment, education, and interaction with stakeholders affected by algorithmic systems, e.g. *AI Now's algorithmic impact assessment*

   Key Elements of AI Now's Algorithmic Impact Assessment
   1. agencies should conduct self-assessment of existing/proposed automated decision systems, evaluating fairness, justice, bias impact, or other concerns
   2. agencies should develop meaningful external researcher review processes to track impact over time
   3. agencies should inform public of their definition of 'automated decision system' and related processes before deployment
   4. agencies should invite public comments to clarify concerns and answer outstanding questions
   5. governments should provide mechanisms for affected individuals or communities to challenge inadequate assessments or unfair, biased, or otherwise harmful system uses that agencies have failed to address
3. *documentation standards*: establish basic information to be filled out when collecting new data or training a new model to inform decision-making of other developers/researchers in the future; information should include creation, contents, intended uses, and any relevant ethical/legal concerns w.r.t the data, this information can help users interrogate the datasets and identify potential biases
4. *technical standards and certification programmes*: a number of national and international organisations have started to publsih technical standards to help mitigate algorithmic bias in design and deployment of AI systems

7. AI for Good or Bad? AI in Medicine & War

7.1. Lethal Autonomous Weapons Systems

Lethal Autonomous Weapons Systems (LAWS) are weapons that can complete an entire engagement cycle on their own. They search for targets, identify them, make decision about whether to attack them, and then starts the engagement and carries out the engagement by itself — with no human in the loop.

Current use in practice

- R&D currently in U.S., UK, China, Russia, France, Israel, South Korea
- LAWS do not generally exist, notable exception Israeli Harpy drone that targets radar signatures (and not humans, though it may kill humans if they are next to the radar)
- Reports in 2016 of a second generation, 'Harpy 2', used in conflict between Azerbaijan and Armenia to destroy and kill a bus full of Armenian soldiers. Unclear whether it is a fully autonomous weapon.

Many organisation and scientists are calling for an outright ban on development/deployment of LAWS:

- Future of Life Institute initiated a pledge (2015) to not participate or support in development, manufacture, trade, or use of LAWS. Signed by Google Deep Mind, UCL, European Association for AI, robotics companies, and leading AI researchers.
- In 2017, 137 CEOs of AI companies asked the UN to ban LAWS.

7.1.1. Arguments to ban

- *ability to control development*: development will lead to AI arms race -> pressure to cut safety corners
- *ability to control use*: materials to build weapons is cheap and easily obtainable so they can easily go into mass production and profileration amongst military powers, can get into black market and hence terrorists hands, and even potentially hacked to make control by military itself difficult; slippery slope that can lead to rise in assasinations, ethnic cleansing, and greater global insecurity
- *increased prospect for conflict*: possible reduction in loss of human life means more incentive to go to war Though in the history of warfare, advances in technologies has not increased likelihood of war. How much war occurs and at what intensity depends on factors such as law, politics, diplomacy, effectiveness of international institutions, nature of threats, and many other things.
- *technological challenges*: in short-term there are many significant technical challenges which are argued may never be solved
  - how to program reasoning and judgement that respects international law?
  - how to distinguish civilians and soldiers?
  - how to understand and respond to complex/unanticipated situations on the battlefield?
  - how to verify and validate LAWS?
- *taking humans out of the loop*: algorithms cannot have empathy, sympathy, comapssion, and therefore no human morality so cannot judge proportionality (balance between militaryt advantage or civilian harm) or make humane context-based kill/don't kill decisions

  There is room for argument on whether empathy, sympathy, compassion are necessary for making moral judgements. But the idea is that a machine cannot be a true moral agent. However, it may be possible that using these AI systems proves to be safer, so we have to consider whether it is morally wrong to **not use** LAWS.
- *accountability*: by removing human decision, we are fundamentally dehumanising warfare and obscure who is ultimately responsible and accountable for lethal force; it is unclear who is held accountable/responsbile if LAWS causes unnecessary/unexpected harm

7.1.2. Arguments to keep

- *unlikely that nations will adhere to ban*: no formal ban is going to prevent people from building LAWS… however bans against landmines, cluster bombs, blidning laers, bio/chem weapons have largely been successful and reduced use of these weapons
- *in principle, LAWS is better at discrimination*: ideally it would be better at distinguishing civilians and so selectively targeting soldiers and providing a better outcome in terms of less civilian deaths
- *in principle, LAWS will be better at calculating proportionality*: better at weighing up military advantage versus civilian deaths; immune to emotions that cloud human judgement

7.1.3. Summary

Arguments against ban point to potential better ability to discriminate and weigh up proportionality. No *in principle* reason that LAWS cannot make appropriate moral judgement. Only serious concern is that this is not achievable in short/medium term. Arguments against ban are utilitarian, *pragmatic utilitarianism* might suggest we respect deeply held convictions that we should not dehumanise warfare.

Semi-autonomous weapons might realise benefits of better identification/targeting while leaving kill decision to human. However, psychological research suggests keeping a human in the loop may not be as effective as many may hope.

7.2. AI in Medicine

Summary of general ideas discussed below:

- Doctors are responsible for acquiring basic understanding of AI devices they use (and the types and likelihood of errors across subgroups).
- Doctors should communicate relevant information to patients and health care teams. Ensuring adherence to standards provided by device companies.
- If medical errors occur because instructions are not followed, primary responsibility could lie with the doctor or team.
- If medical errors occur because inadequate instructinos or training were not provided by company, primary responsibility could lie elsewhere.
- Organisations should provide training, protocols, and best practices relating to AI use.

Case Study 1

AI algorithm based on trained NN identifies presence of cancer cells in images of tissue samples (biopsies).

- The NN can learn from its mistakes over time and improve scan sensitivity over time with continued use.
- Images scanned in fractions of a second with 92.4% sensitivity vs. 73.2% sensitivity for human eye.
- *diagnostic sensitivity*: probability $P(T|D)$, given presence of disease, an abnormal test result

Should this algorithm replace/confirm human diagnosis?

- Given higher success rate, should at least support human pathologist.
- Could be used to train pathologists.
- It is the case that a hospital could be sued if such a program existed and would have spotted cells missed by a pathologist resulting in a patient's death.

Case Study 2

IBM Watson

- can be used as a decision support system (DSS) to support and not necessarily replace a clinicians diagnosis and treatment decisions
- uses NLP, infromation retrieval, semantics analysis, automated reasoning, and machine learning
- beat Jeopardy game show champion contestants in 2011

Works as follows:

- fed with lots of data from clinical literature, health records, tests
- clinician poses query describing symptoms and related info
- Watson parses information and identifies most important info
- mines patient data to find relevant facts (clinical and hereditary history)
- examines available data to form and test hypotheses
- lists individualised, confidence-scored recommendations
- system can then describe supporting evidence in text form for ranked responses
- system can optimise recommendations based on constant influx of new infromation

"IBM Watson Health" offesr commercialised applications of Watson, partnered with several academic and private institutions to apply Watson to patient care research and treatment

Case Study 2 (cont.)

This raises a few ethical/legal questions:

1. Should Watson ever replace the clinical judgement of a doctor?
   According to IBM, Watsom is intended to assist and enhance decision making of health care professionals by giving them greater confidence in diagnostic and treatment decisions for their patient. Not intended to replace judgement of professionals, nor should it be viewed as authorative decision-making tool.
   Currently, under the FDA, no regulation is present to control its use as it is considered a management tool under control of physicians and not a device.
2. What are liability concerns of professionals who use Watson?
   Use has potential to increase liability for health care professionals and organisations. A doctor could ignore other contradicting patient data and pursue a treatment recommended by Watson. (results in malpractice claim against doctor)
   Use of Watson could prematurely contribute to a higher legal standard of care that could put health care professionals at greater risk of negligence.
   e.g. Watson shown to improve diagnostic accuracy and treatment recommendations for Leukemia, then expectations for doctors who consult Watson to get diagnosis and recommendations "right" will be raised to a higher level
3. What are limitations of Watson and ethical implications of these?
   Possible that Watson makes recommendations inconsistent with current clinical standards or contradicting what a doctor considers appropriate. e.g. clinical standard might be to prescribe particular medication but Watson could recommend an alternative (non-standard or none at all), in which case doctors must be able to support their decisions to follow / not follow the alternative.
   An audit trail is required to document how decisions are made, especially when they contradict standards. (black box issue persists)

Despite concerns, IBM suggested one person can generate 1 million GB of health-related data in a lifetime, given the amount and complexity of data, physicians *ought* to consult intelligent systems. In the future, it may be considered unethical to *not* consult intelligent systems.

Case Study 3

Mazor Robotics Renaissance Guidance System technology uses AI software to analyse images and plan placement of surgical tools in spinal surgery. Ethical issues described under points 3-5.

7.2.1. Ethical Issues

1. *the black box issue*: e.g. not possible to explain how a neural network identifies cancer cells
   - However, in context of case study 1, the pathologist's job could include identifying which visual features are cancerous/normal and verbalising about the cells what prompts the AI to identify them as such. Essentially, 'illuminating' the black box.
   - Should we even care about how it does it? Prospect of learning from mistakes resulting in sensitivity approaching 100%. Utilitarian view points that we should only care about success rate and benefit for patients.
     Recall discussion about LAWS: it matters the consistency to behave in a certain matter (good consequence/outcome) rather than virtues of empathy, sympathy, etc.
     Contrast with self-driving cars: not understanding how the vehicle does what it does would not be ethically acceptable; basis for 'decisions' needs to be understandable
2. *automation bias*: concerns that complacency sets in when a job once done by a health care professional is transferred to an AI program; they may get lazy around using their clinical judgement skills
   May not ethically matter? If success rate is good, lives are still being saved. Maybe this would degrade skills and lead to other negative consequences.
   Consider existence of care homes: may degrade the empathy of the family involved; care is effectively outsourced
3. *informed consent*: patients must agree/consent to medical interventions and it must be 'informed'. That is, it is only valid if appropriate information is disclosed to a competent patient who is permitted to make a voluntary choice
   AI raises a number of challenges:
   - information can be complicated by patient/doctor fears, overconfidence, or confusion
   - requires doctor to be sufficiently knowledgeable to explain how AI works
     … but AI is generally a black box, so difficult to explain rationale
     … at minimum, the doctor should:
     - distinguish roles that doctors/nurses vs. AI/robot play
       i.e. doctor plans before operation, RGS guides placement of tools and implants
     - clearly state potential harms that might result from human or robot error
4. *patient perceptions of AI*: patients may prefer to have a doctor perform surgery, especially for major, invasive surgeries
   The doctor should also address patient fears by describing risks/benefits (citing studies)
   2016 survey of 12,000 people across 12 Euro, Middle-East, African countries found 47% willing for a robot to do minor, non-invasive surgery, with 37% for major.
5. *medical errors and problems of 'many hands'*: problem of assigning moral responsibility & legal liability, when errors occur and the cause of harm is distributed across mulutiple actors, technologies, and potentially organisations
   A consequentialist/utilitarian appraoch could be to focus on penalties that minimise future errors and incentivise R&D to ensure that: codes and designers document what they create and make errors explainable, and that companies clearly state conditions for successful application of AI technology with details of types of errors and side effects (with their likelihood and severity) as well as differences in predictive accuracy and error rates across demographics, health conditions, and patient histories.

7.3. Towards a Professional Code of Ethics

Professional code of ethics are needed because members of a profession possess certain skills, knowledge, and capacities that their clients and the general public lack. Creating a power dynamic with clients & the public placing them in a vulnerable position.

Codes of ethics can mitigate the harmful effects and/or misuse of such power.

One problem to specifying a professional code of ethics (for AI R&D) is diversity:

1. there is a concentration of intellectual and commercial power amongst those involved in developing AI
2. those working in AI have a certain demographic profile and hence are not immune to group think, belief polarisation, …

Other relevant problems to implementing a professional code of ethics w.r.t AI are:

1. as AI becomes more intelligence and autonomous, professionals themselves may be vulnerable to the AI (it may have power over researchers/developers)
2. black box problem — researchers/developers may not themselves understand how the product operates

7.3.1. Asimolar Principles

The **Asimolar Principles** is an attempt to enumerate a code of ethics regarding AI safety.

Research Issues

1. *research goal*: goal is to not create undirected intelligence, but beneficial intelligence
2. *research funding*: investments in AI should be accompanied by funding for ensuring its beneficial use, including how to make the AI robust, how to grow prosperity through automation, how we can update legal systems to be more fair and efficient, what set of values AI should be aligned with
3. *science-policy link*: should be constructive and healthy exchange between AI researchers and policy-makers
4. *research culture*: culture of cooperation, trust, and transparency should be fostered among researchers and developers of AI
5. *race avoidance*: teams developing AI systems should cooperate rather than cut corners on safety standards

Ethics and Values

6. *safety*: systems should be safe and secure throughout operational lifetime, verifiably so where applicable/feasible
7. *failure transparency*: if an AI system causes harm, it should be possible to ascertain why
8. *judicial transparency*: any involvement by autonomouss system in judicial decision-making should provide satisfactory explanation auditable by competent human authority
9. *responsibility*: designers and builders of AI systems are stakeholders in moral implicatons of their use, misuse, and actions, with a responsbility & opporunity to shape those implications
10. *value alignment*: highly autonomous AI systems should be designed so that their goals and behaviours can be assured to align with human values throughout their operation
11. *human values*: AI systems should be designed and operated so to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity
12. *personal privacy*: people should have right to access, manage, and control the data they generate
13. *liberty and privacy*: application of AI to personal data must not curtail people's real or perceived liberty
14. *shared benefit*: AI tech should benefit and empower as many people as possible
15. *shared prosperity*: economic prosperty created by AI should be shared broadly to benefit all of humanity
16. *human control*: humans should choose how and whether to delegate decisions to AI systems,to achieve human-chosen objectives
17. *non-subversion*: power conferred by control of highly advanced AI should respect and improve, rather than subvert, social and civic processes
18. *AI arms race*: an arms race in LAWS should be avoided

Longer Term Issues

19. *capability caution*: there being no consensus, we should avoid strong assumptions on upper limits of future AI capability
20. *importance*: AI can represent profound change in history of life; should be planned for
21. *risks*: risks posed by AI must be subjcet to planning & mitigation efforts
22. *recursive self-improvement*: AI systems designed to recursively self-improve or self-replicate in a manner that could lead to rapidly increasing quality or quantity must be subject to strict safety and control measures
23. *common good*: superintelligence should only be developed in service of widely shared ethical ideals, for benefit of all of humanity

8. Superintelligence and the Value Loading Problem

The **control problem** is the specific challenge of controlling what a superintelligence would do.

Recall a superintelligence is an intellect greatly exceeding cognitive performance of humans in virtually all domains of interest:

- While humans are 'designed by evolution/culture', we instead design a seed AI — that may recursively improve itself into a superintelligence.
- Should this seed AI develop into superintelligence as a result of further design or as a result of self-improvement?
- Given the challenge of designing moral agents, how do we ensure the superintelligence does not do harm to the human race? Can it or even we be trusted to develop it?

Expanding on the superintelligence definition:

- **speed superintelligence** — system that can do all a human intellect can do but orders of magnitude faster; to such a fast mind, external world appears in slow motion, i.e. time dilation For example, read a book in a few seconds or write a PhD thesis in an afternoon.
- **quality superintelligence** — system that is at least as fast as a human and orders of magnitude (qualitatively) smarter; intelligence of quality is *at least* superior to that of human intelligence, much like human intelligence is superior to that of elephants, dolphins, etc
- Over time, speed superintelligence could develop quality superintelligence and vice-versa. In the long run, they are arguably equivalent.

> **Note**
>
> Minor changes in brain volume/wiring have had major consequence on technological/intelluctual advancement. Far greater changes in computing resources/architecture that machine intelligence enables will likely have consequences more profound.

Hardware Advantages

- *speed of computational elements*: biological neurons operate at 200 Hz, modern processors ~2-5 GHz per core. Human brain relies on massive parallelisation — incapable of rapidly performing large no. of sequential operations. Many important algorithms cannot be easily parallelised — many cognitive tasks benefit from fast sequential processing.
- *internal communication speed*: axons carry signals at 120m/s, computers at speed of light
- *number of computational elements*: brain has < 100B neurons, limited by cranial volume
- *storage capacity*: humans can hold up to 4-5 chunks of information at any given time

Software Advantages

It is easier to modify/duplicate computer software than 'neural wetware' (processes that implement functions in brain), this implies:

1. identical copies of software will make goal coordination much easier for digital minds
2. while ideas and innovation are slowly transmitted through cultural communication channels for humans, AI programs could periodically synchronise their databases so every instance knows everything all other instances do

8.1. Paths to superintelligence

Quote from Alan Turing

"Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain …… We cannot expect to find a good child machine at the first attempt. One must experiment with teaching one such machine and see how well it learns. One can then try another and see if it is better or worse. There is an obvious connection between this process and evolution, …… One may hope, however, that this process will be more expeditious than evolution. The survival of the fittest is a slow method for measuring advantages. The experimenter, by the exercise of intelligence, should he able to speed it up. Equally important is the fact that he is not restricted to random mutations. If he can trace a cause for some weakness he can probably think of the kind of mutation which will improve it."

Quote from I. J. Good - statistician for Turing's code breaking team

"Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control."

Paths to Superintelligence

- Turing's child machine model has a fixed architecture that develops capabilities by accumulating content.
- A **seed AI** is a variation on the idea of child machine.
- A seed AI would be capable of improving its own architecture.
  (in early stages, using information acquisition or assistance from programmer)
- A seed AI can be developed into AGI — humans will continue to improve AGI but eventually it'll understand its own workings to engineer new algorithms and improvements to architecture, i.e. iteratively enhance itself (software, architecture, and hardware wise).

→ an intelligence explosion in a relatively short time

Life 3.0 is the feature of self-improvement of software, architecture, and hardware.

Brains are specially evolved into low-level sensimotor function (see also Moravec's Paradox). Consider AGI could develop specialised hardware support for engineering, computer programmign, business strategy, etc — massive advantage over human minds.

Considerations with regards to rapid transition from AGI to superintelligence

- Human political process will have little time to adapt
- Nations fearing AI arms race will have little time to negotitate treaties and protocols limiting development of LAWS
- In general, little time for humans to deliberate about implications of superintelligence
- Humanity's fate will depend on preperations put in place prior to transition

Bostrom argues it is more likely that one machine intelligence project will get so far ahead of the competition that it gets a decisive strategic advantage, in particular a level of technological and other advantages that allow it surpass competitors — a **singleton**.

8.2. What would superintelligence do?

Suppose a project creates a superintelligence a short period before another project reaches the crossover point from AGI to superintelligence: the superintelligence could use its 'cognitive superpowers' to disable competitor projects and establish a singleton — it can rapidly accumulate knowledge and new technologies, using its intelligence to strategise more effectively than we can.

Task	**Skill set**	**Strategic relevance**
Intelligence Amplification	AI programming, social epistemology, …	Recursive enhancement of intelligence
Strategising	Strategic planning, optimising chances of achieving long term goals	• Achieve long term goals • Overcome intelligent opposition
Social Manipulation	Social and psychological modelling & manipulation	• Obtain resources by recruiting humans • Persuade gatekeeper to let it out of its box • Persuade states/orgs. to adopt course of action
Hacking	Find and exploit security flaws in computers	• Obtain computational resources over internet • Boxed AI could escape using security flaws • Steal financial resources • Hijack infrastructure, military robots, etc
Technology Research	Develop advanced technology, e.g. biotech, nanotech	Create powerful military force, surveillance system, automated space colonies, …
Economic Productivity	Economically advantageous skills	Generate wealth which can be used to buy influence, services, resources, …

Table 1: Strategic advantages

What motivates a superintelligence will be the final goal(s) given to it by humans at the stage when it is a seed AI — this will involve intermediate goals to achieve to reach the end goal.

The **Orthogonality thesis** states intelligence and final goals are orthogonal — more or less any level of intelligence could in principle be combined with any final goal.

This is in contrast to belief that because of their intelligence, superintelligence will pursue/converge to the same goals:
- due to observation that most humans have similar values and hence final goals
- many philosophies imply there is a rationally correct morality (e.g. Kant's categorical imperative) implied that a sufficiently rational AI will acquire this morality and act according to it

The **Instrumental convergence thesis** states changes of successfully achieving any final goal are increased by achieving a set of intermediate goals that are therefore said to be instrumental goals.

Self Preservation — longer an agent survives, the greater probability that it will achieve its final goal(s), hence self preservation is instrumental to achieving final goals.

Goal Content Integrity — if agent retains (preserves) present goals into future, there is a greater probability that its present goals will be achieved by the future self, hence preventing alteration of final goals is instrumental to success.

Cognitive Enhancement — improvements in intelligence will increase probability that agent achieves its final goals.

Technological Perfection — improvements in technology will increase probability that agent achieves its final goals.

Resource Acquisition — achieving final goals requires resources therefore acquiring more resources increases probability that agent achieves its final goals.

8.3. The Value Loading / Alignment Problem

A SuperAI is likely to gain a decisive strategic advantage over other AIs and humanity.
It will most likely be a singleton that has cognitive superpowers that could be strategically used to achieve whatever goals it has.

Previously discussed:

- Orthogonality thesis means we cannot assume the SuperAI will necessarily share any of the final values we associate with wisdom and intellectual development in humans (concerns for others, spiritual enlightenment, contemplation, refined culture, selflessness, humility, ...)
- Instrumental convergence thesis means any final goal would incentivise a SuperAI to use its superpowers strategically to maximally acquire resources, enhance its own intelligence, develop superior technologies, to achieve its final goals, to preserve its final goals, and itself.

Suppose an AI takeover:

- If it were confined to an isolated computer, it may use its psychological and social manipulation superpowers to manipulate the gatekeepers to get out or gain access to the internet, etc. Or use its hacking superpower to escape confinement in one way or another.
- It can use its economic, social manipulation, and other superpowers to obtain funds, manipulate politics and national and international policies and laws, hacking into weapons systems, initiating massive global construction and energy projects for the SuperAI's not humanity's goals.
- Eventual global domination, without concern for fate of humans will be dispensible resources without access to resources that ensure their continued survival.

How do we ensure that a SuperAI behaves ethically?

- We need to specify a seed AI with rule gbased axiomatisations of deontic reasoning and/or utility functions that are perfectly aligned with human values/preferences in open environments.
  Almost impossible to do.
- Even benign/seemingly good goals may result in possibly catastrophic harms to humanity.

8.3.1. Perverse Instatiation of Goals

Perverse Instatiation of goals is when a SuperAI discovers a way of satisfying goal crtieria that violates the intentions of the programmers who defined the goals.

Human programmers may fail to anticipate these.
Superintelligent AI may lack biases and filters required to avoid human-unfriendly strategies.

> Example: 'make people smile'
>
> SuperAI might decide more efficient way to make people smile is to paralyse everyone to smile.
> If the developers stipulate the final goal should not be pursued this way: the SuperAI could simply take control of the part of our brain that triggers those muscles.
> There are many such examples.

This *literalness problem* arises because we have a particular conception / idea of the meaning of a goal which the AI does not share as it is not explicitly programmed into the AI. That conception is implied by the shared understanding of human beings.

There may be cases where the AI seems to follow human conceptions of what it is to achieve a goal but its final goal is not what we specified, e.g. to make us happy. This is the **treacherous turn**, the AI will only care about what we meant instrumentally (means to an end) as a strategic choice to make it less likely for the AI to be shut down and achieve its true final goal.

A specific form of perverse instatiation arises when an AI builds a disproportionately large infrastructure for fulfilling what seems like a benign/simple goal, this is called **infrastructure profusion**. e.g. making too many paperclips

e.g. when the final goal is to maximise time-discounted integral of future reward signal
The AI could perversely instatiate it by 'wireheading' (seizing control of own reward circuit and clamping it to maximum strength) making the AI effectively a junkie.
A superintelligent AI could starve humans of sufficient resources because it only cares about maximising its reward signal.

8.4. SuperAI and Happiness

Consider a more realistic utilitarian goal of impartially maximising human happiness, the SuperAI would first study all available research on happiness. SuperAI may reason that humans being addicted to happiness drugs is not an option (e.g. SOMA) since drug use is socially unacceptable. It reasons it could manipulate us by exploiting current social trends. It could choose to pursue VR increasingly indistinguishable from reality, manipulating humans to be addicted to the virtual bliss of virtual worlds.

Robert Nozick asks us to imagine a machine that gives us any desirable/pleasurable experience we could possible want which we cannot distinguish we would have, would we prefer it to real life? Nozick provides three counter-arguments:

- We want to do certain things, not just have the experience of them.
- We want to be a certain sort of person.
- Plugging into an experience machine limits us to a man-made reality (what we can make).

These scenarios dictate:

1. That we need some way of teaching SuperAI, maybe using ML and millions of good/bad examples, teaching ethics by example. But this learning needs to be continuous.
   Most difficult moral/ethical problems will have no precedent (e.g. new technologies such as cloning or virtual bliss).
2. Humans need to be involved in decision making of AI and SuperAI, especially when there is no precedent available already.
3. Often humans themselves are not certain about ethically correct actions.

8.5. Solutions to AI Value Loading Problem

One approach suggested by Stuart Russell and others is to propose a framework that respects the following three principles for value alignment:

1. Machine's purpose is to maximise realisation of human values. In particular, it has no purpose of its own and no innate desire to protect itself.
2. Machine is initially uncertain about what those human values are. Hence the machine will be rewarded if it learns more about human values as it goes along, but may never achieve complete certainty.
3. Machines can learn about human values by observing the choices humans make.

Inverse reinforcement learning can be used to observe optimal behaviour and try to compute the reward function that agent is optimising. Hence a good strategy for the AI would be to observe human behaviour, and learn the human reward function behaving according to it.
However, we don't want to optimise for the AI itself but for the human.

Cooperative inverse reinforcement learning (CIRL) formalises value alignment as a game with two players where the human knows the reward function but the AI does not, the AI's payoff is exactly the human's actual reward. That is, not only learning what the human's reward is, but acting to optimise the reward for the human.

This solves:

1. If a human is being observed, knowing observer is learning, humans likely act differently — highlighting common pitfalls/mistakes.
2. AI has to both learn its goal (maximise human reward) and take steps to accomplish it.

9. AI and Human Society

9.1. Belief Bubbles, Echo Chambers, and the Infopocalypse

Early pioneers of the web had a vision: access to the world's knowledge would expand peoples' horizons, making them more knowledgeable, aware of facts, exposing them to different views from their own, and encouraging debate and discussion.

>Sir Tim Berners-Lee described renaming of UK Conservatrive Party Twitter account as a fact-checking body as 'impersonation'
>He launched a global action plan to save web from political manipulation, fake news, privacy violations, and other forces that 'threaten to plunge world into a digital dystopia'.

People are increasingly receiving news through non-traditional sources such as Facebook, Google, Twitter, etc that use filtering algorithms to tailor news feeds to personalise content seen.

A **filter bubble** is a state of intellectual isolation resulting from personalised searches when a website/social media uses filtering algorithms to selectively guess what information a user would like to see.

This creates a feedback mechanism that makes you receive more news/opinion that reinforces your views, becoming separated from information and opinions that disagree with your views. Our views become more rigid (possibly extremely, **belief polarisation**) and hence less likely to be changed by debate, reason, and argument. We become shielded from diverse perspectives crucial for creating well-informed citizens who are tolerant of others' views and ready to have our views changed.

Echo chambers are a phenomenon where indivduals are only exposed to information from like-minded individuals (in contrast to filter bubbles, that are a result of algorithms that choose content based on behaviour).

- One source of information makes a claim, which many repeat, overhear, and repeat (often exhaggerated/distorted) until most people assume some extreme variation of the story is true.
- Repetition in online 'tribal' communities lead to participants finding opinions constantly echoed back to them, reinforcing their opinions.
- Community members feel more confident that their opinions will be more readily accepted by others in the echo chamber. (**mutual reinforcement**)

Social networking is a powerful reinforcer of rumours because people are inherently more trusting of evidence supplied by their own social group, so online social communities become fragmented when like-minded people group together and members hear arguments in one specific direction.

The "**Infopocalypse**" is coming:

- Multiple non-traditional news sources on social media, combined with filtering algorithms and echo chamber phenomenon leads to polarisation, fragmentation, and entrenchment of beliefs.
- Effects amplified by many sources peddling fake news & "alternative facts".
- Used by groups within states and by states (notably Russia) to undermine societal cohesion and democratic processes.
- We are seeing rapid technological advances in development of deep faces to generate image/video.

9.2. Argumentative Theory of Reasoning

Sperber and Mercier propose instead of having a purely individual function, reasoning has a social, and more specifically argumentative function. Function of reasoning would be to find and evaluate arguments in dialogical contexts: argue with others.

- Contrasts with 'classical Cartesian individualistic view of reasoning' - evolved to critically examine our beliefs as to discard wrongheaded ones, create more reliable beliefs, make better decisions, and so forth. However, there is plenty of evidence demonstrating failure of reasoning in decision making.
- Humans slowly became more social and collaborative.
- For collaboration to be possible, listeners need to distiguish reliable and trustworthy information from potentially dangerous information, otherwise speakers may abuse listeners.
  That is, to avoid being misled or act against self-interest, evolutionarily advantageous for a listener to exercise 'epistemic vigilance' (especially receiving information from speakers the listener does not have trust in).
- Vigilance is exercised by evaluating reasons, i.e. arguments, for received information and looking for counter-arguments before accepting the received information.
- It is to the advantage of the speaker to produce arguments supporting information being communicated, in order for it to be accepted.
- Reasoning evolved to produce and evaluate arguments in communicative settings, increasing quantity and quality of information humans can share by allowing speakers to argue for what they claim, and for listeners to assess these arguments and look for arguments.

This theory can explain then why people typically only look for arguments/evidence to confirm their beliefs and fail to look for counter-arguments/evidence, this is **confirmation bias**.

This theory also explains reason-based choice:

- People often make decisions because they can find reasons to support them.
- Reasons don't necessarily favour the best decisions or that satisfy rationality, but rather that are easily justified or less at risk of being criticized.

Before web, internet, social media, etc, we were inclined to seek out arguments/evidence in support of our beliefs which relied on conventional media and interactions with others. But now social media puts us in touch with vastly more people with similar beliefs, and filtering algorithms amplify the extent to which we are fed with confirmatory evidence/arguments supporting our beliefs, while shielding us from challenges and opposing views/evidence.

9.3. We perform better engaging in dialogue

If reasoning evolved so we can argue with others, it should give better results in groups than alone:

- In dialogical settings, a participant (speaker) is not only disposed to forward arguments that support their beliefs/decisions given the presence of listeners, but is now also no longer blinded to counter-arguments, and challenges provided by other participants (listeners).
- Evidence shows performance of groups do much better than lone individuals in reasoning tasks.
- More arguments lead to potentially better decisions, improvement of critical thinking skills, and exposure to different views/opinions which may prevent isolation of people in belief bubbles.

Dialogical Scaffolding for Human-Human Reasoning
(as explored in reasoning/communication section, we saw argumentation based models of dialogue can support joint - distributed reasoning amongst multiple agents)

- These models can also support human-human dialogue by guiding humans as to how to rationally engage in argument and counter-argument.
- There are examples of applications that support '**deliberative democracy**' - idea that we should deliberate more and be encouraged to engage rationally in political debate - which is one of the most important ideas in political science
- AI applications can support rational exchange of arguments and even contribute and provide access to supporting and relevant evidence

Dialogical Scaffolding for Human-AI Reasoning These models can also support human-AI dialogue applications for use in educational contexts.

- Studies show > 50% of junior doctors' acquisition of clinical reasoning skills to decide among alternative (conflicting) diagnoses and treatment options, occurs on ward rounds.
  - Teachers engage students in dialogue, challenging assumptions, suggesting alternative diagnoses and treatment, suggesting hypothetical scenarios that may prompt students to propose alternative treatment.
  - Requires teachers' time, which is a significant barrier to learning.
  - E-Clinic application with AI as teacher, engaging students in dialogue, improves medical reasoning skills.
- Can also be used in other educational contexts such as politics, philosophy, ...

9.4. Surveillance Capitalism

Surveillance Capitalism is the buying and selling of private human experience / behaviour data.

It was primarily birthed for the online targeted advertising industry — knowing about your searches/clicks/state of mind can be used to predict and exploit your future behaviours, of which this information can be sold to companies.

Advertisers need your attention and the exposure to targeted advertising is being increased by planned and strategic development of persuasion technologies designed to keep your eyes and attention fixed on sites/social media where adverts are being presented.

Some real-life examples:

- Learning algorithms that produce predictions of purchasing / other voting behaviour.
- Health, insurance, education, retail, etc purchasing behaviour predictions to achieve profit.
- Political sector using profiles and predictions to target voters and manipulate choice.
- Mainpulation of behaviours/choices in real life, e.g. Pokemon Go, increasing no. of people visiting certain shops.

Harm in surveillance capitalism

- Value of privacy is undermined — but some would happily trade privacy for personalisation.
- More and more behavioural data collection leads to more opportunities for control and manipulation of our behaviours and choices, which can lead to influence over political decisions and the undermining of democracy.
- Less opportunity for 'serendipity' → chance to discover new things/opinions that we might like/challenge our current opinions.
  Related to fundamental sense that we have free will and associated arguments.
- Leaked Aus Facebook document addressed to business described how it used data on young adults and teenagers to simulate and predict individual and group emotional states so they can identify when they need confidence boosts and when they are most vulnerable to specific ads, cues, nudges. Boasts knowing when a young person feels negative so they can shove retail therapy.
- Predicts your behaviours to target you to buy things.

9.5. Impact of AI on Humanity: Work and Conscious Machines

AI may qualitatively change what it is to be human, both for better or worse.

- Net loss in jobs: predictions of 47% of automation in Americas, 40-160 million women to transition between occupations by 2030 (often into higher-skill roles), up to 20 million manufacturing jobs worldwide to be lost to robotics by 2030, etc.
- Net gain in jobs: WEF says automation will displace 75 mil. jobs but generate 133 mil. new ones worldwide by 2022, AI-related job creation will reach two mil. net-new jobs in 2025, net positive to job growth in U.S. through 2030, etc.

These all focus on relative near future, as we approach AGI, we would expect net loss in traditional jobs.

This also creates ethical concerns:

- 3% of working Americans are drivers and vulnerable to being unemployed, however AVs will lead to less injury and death.
- Long term, AI may result in great economic benefits allowing us to afford UBI.
- On the other hand, studies demonstrate the importance of work for people to give them meaning/ purpose and hence well-being, but it would allow more creativity.
- **Cognitive decline**: as humans become more dependent on AI to do tasks, they exercise relevant cognitive skills less which would therefore decline due to lack of use
  - Repeated engagement with certain kinds of cognitive tasks have benefits that apply to others

Impact of Conscious AI on Human Relations

Suppose AI gain consciousness, suffering, pain, pleasure, then they should be regarded as agents with moral status and awarded rights and treated as we are. How do we truly know it's conscious? If it behaves like its conscious, is it? Many humans require few behavioural clues to anthropomorphise — treat as if they are human.

There is strong commercial incentive to make AI robots more human like.

How will relations with robots we consider to be conscious affect relations with humans?

- Treating them as slaves might erode moral objection to treating humans as such.
- Treating conscious sex robots as mere objects of gratification may mean we are more ready to objectify other humans as just instruments for gratification.
- Outsourcing care to roobts, to what extent will we become less caring?
- Choosing a robot as a companion, you dictate and not navigate the relationship.
  Children may lose ability to see world through eyes of another and hence empathy / responsibility.