This document is intended to be exported as Anki flashcards.
Answers generated/derived by/from AI are coloured in blue! (Gemini/NotebookLM in use)
"Open-book" answers are marked as purple, corrections from model solutions are coloured as green.
Any arguments arising from group discussion are marked as fuchsia.

7CCSMEAI
Practice Questions

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

Contents

1. Week 2: Intelligence

Intelligence could be said to be "substrate neutral" if:

1. ✗ 'Real' understanding is considered as being logically sufficient for Intelligence and real understanding is substrate neutral.
   This implies there may be entities that are intelligent that do not have real understanding.
2. ✓ Intelligence is defined in terms of information and computation, and information storage and computation are substrate neutral.
3. ✗ Intelligence is said to measure an ability to achieve complex goals in a wide range of environments, independently of the environment's substrate.
   Substrate of environment has nothing to do with ability to achieve complex goals.
4. ✓ Intelligence is defined as the ability to achieve complex goals in a wide variety of environments, and this ability is substrate neutral.
5. ✗ Intelligence is defined in terms of information and computation, and is a measure of the ability to achieve complex goals in a wide range of environments, independently of the environment's substrate.
   Substrate of environment has nothing to do with definition of intelligence.

Which of the following correctly describes the Impossibility Reply to the Chinese Room Argument?

1. ✗ The Impossibility Reply concedes that Searle is right when he argues that it is impossible for the system in the room to understand language and know meaning of words. But if let loose in the world (e.g. as a robot) the system would come to understand meaning by interacting with the world and with others.
2. ✓ The Impossibility Reply challenges the argument's assumption that the symbol manipulating program P, simulating Chinese understanding, is possible given the laws of physics.
3. ✓ The Impossibility Reply challenges the argument's assumption that the symbol manipulating program P, simulating Chinese understanding, by claiming that P is not a nomic possibility.
4. ✗ The Impossibility Reply challenges the argument's assumption that any actual digital computer C will not differ from P in any sense that makes it more intelligent than P, by arguing that it is not possible to compare digital computers with the program P.
5. ✗ None of the other options correctly describe the Impossibility Reply.

2. Week 3: Consciousness

A feature of human consciousness is awareness of one's own mental States (meta-awareness).
Which of the following are correct statements about this feature of meta-awareness?

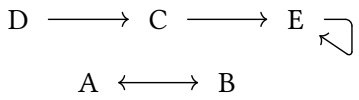
1. ✗ If meta-awareness is understood as having information about informational states, then meta-awareness cannot be implemented in an AI.
2. ✗ If meta-awareness is understood as having information about informational states, then there must be an accompanying qualia: what it feels like to have information about informational states.
3. ✓ Meta-awareness enables social coordination, as being aware of our own beliefs, goals, desires, perceptions enables us to have a better understanding of others' mental states.
4. ✗ Meta-awareness restricts the flexibility needed to respond to novel situations, because by directing attention to a given mental state, we ignore other mental states.

A philosopher argues that having conscious states (in the sense that there is phenomenological experience/qualia) is a logically necessary condition for AGI. **This means that:**

1. ✓ A machine may be conscious but not accomplish any goal at least well as a human.
2. ✓ It is highly likely that the philosopher thinks that epiphenomenalism is false.
   a philosophical position that mental events are caused by physical events in the brain, but do not cause physical events in return
3. ✗ It is highly likely that the philosopher thinks that epiphenomenalism is true.
4. ✗ Given the philosopher's claim, it cannot be the case that the philosopher takes a functionalist view of consciousness.
5. ✗ Given the philosopher's claim, it cannot be the case that the philosopher takes a physicalist view of consciousness.
6. ✗ None of the other statements follow from the philosopher's claim.

3. Week 4: Reasoning and Communication

Consider the argument framework below. Which of the following statements are true.



1. ✗ The grounded extension is the empty set.
2. ✓ The preferred extensions are $\{D, A\}$ and $\{D, B\}$
3. ✗ The preferred extensions are $\{D, A, E\}$ and $\{D, B, E\}$
4. ✓ The grounded extension is $\{D\}$.

Which of the following are correct statements about the 'symbol grounding problem'?

1. ✗ None of the other statements about the symbol grounding problem are correct.
2. ✓ The symbol grounding problem is to some extent solved by machine learning, in the sense that its representations of the world are extracted from statistical data capturing the rich statistics of the real world.
3. ✓ The robot reply to Searle's Chinese Room argument, illustrates how the symbol grounding problem can be solved.
4. ✗ The symbol grounding problem is primarily associated with machine learning based approaches to Artificial Intelligence, as machine learning is primarily sub-symbolic (statistical).
5. ✓ The symbol grounding problem is primarily associated with logic- based approaches to Artificial Intelligence that require that the symbolic elements of a representation are hand coded by the designer of the system rather than being grounded in data from the real world.

4. Week 5: Ethics and Morality

Suppose an autonomous spacecraft called SpaceVac is launched into space with the mission to identify 'space debris' (the many dead satellites and fragments of man-made machinery discarded in Earth orbit, which if left unchecked could pose significant problems for future generations by making access to space increasingly difficult, or at worst, impossible). SpaceVac is continually updated with a list of space machinery that is currently in use (and so is not considered as space debris). SpaceVac has sensors that are able to detect and collect (with 100% accuracy) for later disposal, only man made machinery in space that is not on this list; that is to say, only machinery that is identified as space debris.

According to Moor's categorisation of ethical agents, SpaceVac would be categorised as:

1. ✓ An implicit ethical agent.
2. ✗ A full ethical agent
3. ✓ An agent with ethical impact.
4. ✗ Both an explicit ethical agent and an agent with ethical impact.
4. ✗ An explicit ethical agent

An artificial moral agent is programmed so as to be an ideal Kantian agent that implements Kant's categorical imperative. **The agent is said to be:**

1. ✗ A top down implementation of a deontic agent that judges whether to follow any given rule for how to act, by assessing whether the consequences of following the rule would lead to more harm than good (as measured by some utility function).
2. ✗ A top down implementation of a consequentialist agent that judges whether to follow any given rule for how to act, by assessing whether its own goals would be blocked if other agents also followed the given rule.
3. ✓ A top down implementation of a deontic agent that judges whether to follow any given rule for how to act, by assessing whether its own goals would be blocked if other agents did not follow the given rule.
4. ✗ A top down implementation of a deontic agent that judges whether to follow any given rule for how to act, by assessing whether the goals of other agents would be blocked if the agent followed the given rule.
5. ✓ A top down implementation of a deontic agent that decides to act according to a given rule in a given situation, having made a calculation that it can achieve its own goals, if all other agents also chose to follow the same rule when faced with the same situation.

5. Week 6. Algorithms

Typically, an agent is said to be morally responsible for a criminal action if and only if:

1. ✗ The agent understood the ethical impact of the action, and was responsible in the sense that if it had chosen otherwise it would have been rewarded.
2. ✗ Punishing the agent will ensure that the system does not choose the criminal action in the future.
3. ✗ The agent understood the ethical impact of the action, and the decision procedure that led to choosing the action was faulty.
4. ✗ The agent's decision procedure that led to choosing the action was faulty.
5. ✓ The agent understood the ethical impact of the action, and freely chose to perform the action

A sophisticated AI system (FinAd) has been specifically trained to advise people on how to invest their savings. However, FinAd has been hacked and its code corrupted by a criminal organisation. As a result, a user who follows the advice of FinAd is defrauded of all her life savings. A judge is asked to decide who or what is criminally liable.

Given existing legislation relating to liablity, which of the following are judgements that could be made?

1. ✗ The judge might declare that the designers of FinAd are liable under the perpetrator via another doctrine, because the way they designed and programmed the AI system and its training protocols, meant that it was vulnerable to being hacked and corrupted.
2. ✓ The judge might declare that the designers of FinAd are liable under the natural probable consequence doctrine, because the designers did not put in place levels of security that would ensure that FinAd could not hacked.
3. ✗ The judge could argue that FinAd is a service, and so the rules of negligence apply. In this particular case, the judge recommends that the designers of FinAd should therefore be prosecuted and that the prosecution needs to prove that the designers of FinAd are morally responsible as this term is currently understood in philosophy and law.
4. ✓ The judge could argue that FinAd is a service, and so the rules of negligence apply. In this particular case, the judge recommends that the designers of FinAd should therefore be prosecuted and that the prosecution needs to prove that the designers of FinAd had a duty of care that they breached, and as result of the breach, harm was caused to the user.

6. Week 7. AI for Good or Bad?

Which of the following arguments might be made by a utilitarian who takes a pragmatic approach to thinking about the morality of using AI in War?

1. ✓ If empathy and compassion are neccessary in order to ensure that use of weapons in war minimises harm, and AI systems cannot in principle possess the virtues of empathy and compassion, then we should ban lethal fully autonomous weapons (LAWS).
2. ✗ If empathy and compassion are neccessary in order to ensure that use of weapons in war minimises harm, and AI systems cannot in principle possess the virtues of empathy and compassion, then this does not mean we should ban lethal autonomous weapons (LAWS).
3. ✓ We should respect deeply held convictions that warfare should not be dehumanised, and so even if the use of AI in fully autonomous weapons means better identification of legitimate targets, better targeting and better calculations of proportionality, we should always ensure that humans are in the loop when making kill decisions.
4. ✓ If it is highly probable that the use of LAWS will mean an increase in warfare and so lead to more death an injury as compared with continued use of conventional weapons that do not use AI, then this is a stong argument for banning LAWS
5. ✗ None of the other options describe arguments that would be made by a pragmatic utilitarian.

In 2030, a medical application - TREAT-AI - is used to make treatment recommendations. Suppose that in 2030 the law of criminal liability applies to the use of applications such as TREAT-AI. Suppose also that TREAT-AI was only trained on clinical treatment guidelines written in English, so that when TREAT-AI was used in India, the English treatment recommendations are then automatically translated by AI, from English into Hindi. Moreover, those responsible for approving the use of TREAT-AI in India were aware that the translation of certain medical terms, from English to Hindi, can result in changes in the meaning of these terms.

Suppose then that TREAT-AI is used to make a treatment recommendation in India, but because of a translation error, a patient is injured after being treated according to the a TREAT-AI recommendation. It has been established, to the judge's satisfaction, that those responsible for approving the use of TREAT-AI in India did not intend that the use of TREAT-AI in India would result in harming patients.

Which of the following judgements could well be made by a judge trained in the legal uses of AI and AI bias?

1. ✗ Those responsible for approving the use of TREAT-AI in India could be charged under the perpetrator via another doctrine.
2. ✓ Those responsible for approving the use of TREAT-AI in India could be charged under the natural probable consequence doctrine.
3. ✗ Those responsible for approving the use of TREAT-AI in India could be charged under the direct liablity doctrine.
4. ✗ The use of TREAT-AI in India is an example illustrating feedback bias.
5. ✓ The use of TREAT-AI in India is an example of data bias relative to statistical standards.
   Since TREAT-AI was only trained on clinical treatment guidelines written in English, this is clearly a case of data bias relative to statistical standards.

7. Week 8: Superintelligence and the Value Loading Problem

Dr. Ligdom and Dr. Selim debate whether a future superintelligent AI that they call HAL, could cause significant harm to human beings. At the outset of their discussion, they agree on the following assumptions:

1. HAL would be switched on and start to operate and act in 2060.
2. HAL is not the only superAI (HAL is not a singularity).
3. HAL would have an off button that humans could use to switch HAL off. Also, HAL would be given a final goal of shutting itself down on a precise date in 2070.
4. Although the environment E in which HAL starts to operate in 2060 is a relatively open environment, it would possible to identify all the relevant facts that describe the environment E when HAL starts to operate in 2060, and predict all the decisions HAL would be faced with at the moment HAL starts to operate.

Which of the following are arguments supported by researchers who have studied the issues of intelligence, morally good actions and superintelligent AI:

1. ✓ Dr. Selim finds Kant's theory of morality very appealing, and so he argues that the more rational and intelligent that one is, the greater the probability that one would choose a morally good action, and so the more intelligent an AI is, the less likely it is to cause significant harm to human beings.
2. ✗ Dr. Selim argues that before switching on HAL, one could use another superintelligent AI – who she calls SOCRATES to gather all the facts about E and predict all the decisions HAL would be faced with in the environment E. Before being switched on, HAL could then be rewarded by SOCRATES for making the morally right choices for all these potential decisions, and this would then guarantee that HAL would not cause harm throughout the lifetime of its operation.
3. ✓ Dr. Ligdom argues that HAL could manipulate humans so that they do not switch off HAL if things start to go wrong and HAL starts to cause harm. This is because of the instrumental goal of self-preservation.
4. ✓ Dr. Ligdom argues that the instrumental goal of self-preservation would always imply that a Superintelligence pursues the anthropomorphic goal; that is to say, HAL will manipulate humans into thinking that HAL is alive, conscious and experiences feelings. This means that the anthropomorphic goal can also be considered to be an instrumental goal.

8. Week 9: AI and Human Society

Which of the following correctly describe aspects of Surveillance Capitalism?

1. ✗ Surveillance Capitalism describes how data recording our online behaviour is used to train autonomous machine learning systems without our consent.
2. ✓ Surveillance Capitalism describes how data recording our online behaviour is used to predict what we desire and then use these predictions for targeted advertising.
3. ✓ A key reason for why behavioural predictions is useful, is because of the attention economy business model used by social media.
4. ✗ Surveillance Capitalism describes how increased profits can be obtained by companies who use AI to monitor the activities of their employees.
5. ✓ Surveillance Capitalism describes how our online behaviour could be used to create behavioural profiles of users, and how this information could be used to manipulate the political beliefs and choices of users.

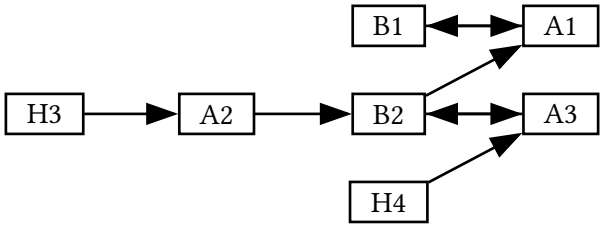
Which of the following statements are correct?

1. ✗ Social media filtering algorithms amplify the confirmation bias of social media users by exposing them to opinions and views that challenge their beliefs.
2. ✗ The phenomenon of echo chambers on social media is primarily because of the use of filtering algorithms that selectively reinforce the beliefs and opinions of social media users.
3. ✗ The argumentative theory of how reasoning evolved explains why we make better decisions when we are required to communicate the reasons for our decisions to others.
   incorrect as the theory predicts that we make decisions based on the ease with which reasons can be communicated rather than because they are optimal
4. ✗ AI applications that support dialogue can potentially lead to an improvement in critical reasoning skills because the users of these applications can be exposed to a wider range of reasons and arguments that support the users' beliefs and opinions.
5. ✓ None of the other statements are correct.

9. Assorted MCQ (extras & 2024-25)

Which of the following describes Moravec's Paradox and the reasons for the paradox?

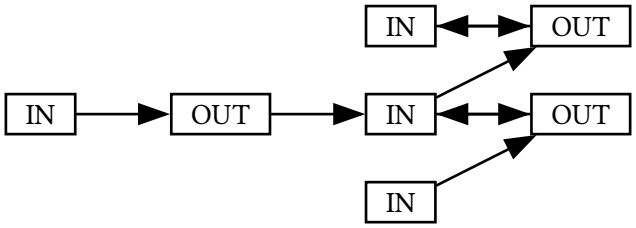
1. ✗ High level reasoning tasks seem easy to humans, whereas computers require huge computational resources to implement these tasks. This is because human brains dedicate massive amounts of neural wetware for implementing these tasks.
2. ✗ High level reasoning tasks seem easy to humans, whereas computers require huge computational resources to implement these tasks. This is because of the problem of the intractability of common-sense reasoning.
3. ✓ Low level sensorimotor tasks seem easy to humans, whereas computers require huge computational resources to implement these tasks. This is because human brains dedicate massive amounts of `neural hardware` to implementing these tasks.
4. ✗ Low level sensorimotor tasks seem easy to humans, whereas computers require huge computational resources to implement these tasks. This is because human brains dedicate massive amounts of `neural wetware` (the processes that implement functions in the human brain) to implementing these tasks.



Given the argument framework above, **which statements are true?**

1. ✓ Grounded extension is $\{B_1, B_2, H_3, H_4\}$
2. ✗ Grounded extension is $\{B_2, H_3, H_4\}$
3. ✗ Grounded extension is $\{H_3, H_4\}$
4. ✗ None of the other statements are true
5. ✗ There are two preferred extensions
6. ✓ There is one preferred extension

Only possible labeling:



The Turing test falls into:

- ✓ Human-based and reasoning-based: systems that think like humans
- ✗ Human-based and behaviour-based: systems that act like humans
- ✗ Ideal rationality and reasoning-based: systems that think rationally
- ✗ Ideal rationality and behaviour-based: systems that act rationally

Professor Ligdom proposes a test that can be used to determine whether a computer is or is not conscious. He proposes that a panel of psychologists, neuroscientists, AI specialists, and researchers in consciousnesss studies should decide whether a computer is conscious. He argues that if:

1) After two hours of questioning, the computer's responses lead all panel members to make a judgement that there is a greater than 50% chance that the computer is conscious, and

2) The computer's architecture is designed so that there there are a number of different modules designed for specific tasks, and there is a central module that broadcasts information that it is accessible to all these modules, and that there is feedback from these modules to the central broadcasting module.

then the computer should be considered as conscious.

Which of the following statements are correct?

1. ✗ Given the two criteria for judging whether a computer is consious, Professor Ligdom commits to a functionalist view of consciousness.
2. ✓ Given the two criteria for judging whether a computer is consious, it is highly likely that Professor Ligdom thinks that consciousness is substrate neutral.
3. ✗ Given the two criteria for judging whether a computer is consious, Professor Ligdom is committed to the view that AGI (artificial general intelligence) is a neccessary condition for consciousness.
4. ✓ Given the two criteria for judging whether a computer is conscious, Professor Ligdom ***is not*** committed to the view that AGI (artificial general intelligence) is a neccessary condition for consciousness.
5. ✓ Given the two criteria for judging whether a computer is consious, the panel of judges may include members who are commited to ***property dualist*** theories of consciousness (recall that property dualism is the idea that there are two distinct and disjoint kinds of properties: physical ones and conscious ones)

Which of the following correctly describe problems with top down implementations of deontic autonomous vehicles that are expected to act ethically in complex and unpredictable environments?

1. ✗ These vehicles are trained on data that is used to describe a wide range of ethical scenarios, and the appropriate ethical choices that should be made in these scenarios.
2. ✗ If data specifying appropriate ethical choices (in all the scenarios that the designers anticipate) is used to train these autonomous vehicles, then this will guarantee that these vehicles will always make ethical choices that are aligned with the values of the designers.
3. ✗ Although there may be specific ethical scenarios in which the vehicles' deontic rules do not apply, current approaches to transfer learning can be used to revise the rules so as to ensure they do apply to novel and unexpected scenarios.
4. ✓ All the other options are incorrect.

When AI algorithms are used in narrow and predictable contexts, a basic requirement is *procedural regularity*. Which of the following statements about procedural regularity are correct?

1. ✗ If one can ensure that the decision making process of algorithms are transparent, then this means that the requirement of procedural regularity has been met.
2. ✓ Tools for procedural regularity can provide assurances that each decision made by an algorithm is reproducible from the specified decision policy and the input for that decision.
3. ✓ Tools for procedural regularity can provide assurances that if a decision requires any randomly chosen inputs, those inputs are beyond the control of any interested party.
4. ✓ Cryptographic commitments can be used to ensure procedural regularity.
5. ✗ Procedural regularity is proposed as a technique to solve the black box problem. This technique requires that experts explain why they believe a machine learning algorithm makes a decision, and then uses these explanations to train the machine learning algorithm to generate accompanying explanations for future decisions.

What does it mean to say that a superintelligent machine 'perversely instantiates a goal'?

1. ✗ A superintelligent machine satisfies a goal by manipulating humans into achieving the goal.
2. ✗ A superintelligent machine fails to satisfy its final goal, and instead achieves a goal that is misaligned with the values of the programmers who defined the final goal.
3. ✓ A superintelligent machine discovers a way of satisfying the criteria of its final goals that violates the intentions of the programmers who defined the goals.
4. ✗ A superintelligent machine fails to satisfy its final goal, and instead achieves only an instrumental goal.
5. ✗ A superintelligent machine fails to satisfy its final goal, since it recognises that achievement of the final goal would violate the intentions of the programmers who defined the final goal.

When considering the ethical dimensions of using AI in Healthcare, the "many hands" problem is the problem of assigning moral responsibility, and perhaps legal liability, when errors occur and the cause of the harm is distributed amongst multiple actors, technologies, and perhaps organisations. Which of the following statements are correct?

1. ✗ The 'black box' issue (that is especially relevant to machine learning systems) helps solve the many hands problem, as all decisions can be recorded and subsequently examined to make an assessment as to who or what bears responsibility/is legally liable (in a manner similar to the way the black box on aeroplanes, also known as flight recorders, records important data that helps investigators determine the cause of a crash and prevent similar incidents).
2. ✗ A consequentialist approach to addressing the many hands problem would be to exclusively penalise the companies who developed the AI systems whose use resulted in medical errors. This will incentivise more rigorous safety testing procedures and so reduce the possibilty of future errors.
3. ✗ Doctors who use these AI systems, cannot be held responsible, as they are merely 'end-users' of these systems.
4. ✓ None of the other options are correct.

10. Essay Q: Chinese Room Argument

Describe the Chinese Room that the philosopher John Searle uses to then develop his Chinese Room Argument (5 marks)

Suppose you have a room in which a person/machine/etc is placed, questions written in Chinese are fed into the room, and the room is expected to produce answers back in Chinese.

A person who doesn't know Chinese is then locked in the room with a set of boxes of Chinese symbols (a database) and instructions on how to convert (a program) a given set of symbols (representing the question) into another set of symbols (representing the answer), thus if a person uses this set of instructions they can appear as if they know Chinese to an outsider who isn't aware of how the inner workings of the room. This enables the person in the room to pass the Turing Test for understanding Chinese yet they do not understand a single word of Chinese!

List the three assumptions that Searle uses to conclude that digital computers cannot have human level intelligence (9 marks)

1. Symbol manipulating program P is a kind of digital computation that performs this symbol transformation and is assumed to exist in this world
1. The program does not possess understanding hence is not intelligent
2. A digital computer C will not differ from P in any sense that makes it more intelligent than P

So $\therefore$ if P is not intelligent, then any actual digital computer C cannot be intelligent.

Searle demonstrates that a computer can only manipulate symbols, but it cannot understand their true meaning, which is one of the hallmarks of intelligence.

How can the AIXI definition of intelligence be used as an objection to the conclusion of Searle's Chinese Room argument? (12 marks)

AIXI defines intelligence as a mathematically defined measure of the ability of an entity to achieve complex tasks, which explicitly does not require an understanding of the complex tasks at hand.

This is an objection to the assumption that "the program does not possess understanding and hence is not intelligent", as according to the AIXI definition:

- it does not need to possess understanding in order to be intelligent
- the task of answering questions in Chinese is considered as achieving the complex goal of conversing in Chinese hence under AIXI, can be considered intelligent

Since it attacks one of the key assumptions of the conclusion of Searle's argument, it is also an objection to the conclusion as a whole.

The conclusion of the Chinese Room argument is that any actual digital computer cannot be intelligent. A key assumption is that no real understanding is present in the Chinese Room, and that understanding is a neccessary condition for intelligence (if there is no understanding then this implies that there is no human level intelligence). However, the AIXI definition of intelligence - Intelligence measures an agent's ability to achieve goals in a wide range of environments – does not imply that "understanding" is a neccessary condition for intelligence.

How might an argument based on functionalism be used to object to Searle's claim that there is no real understanding? (8 marks)

According to functionalism, If a system S1 produces the same input/output pairs as a system S2 that we know to possess real understanding, then SI possesses real understanding. Now, the Chinese room, by assumption, generates all the linguistic responses one would expect of a system that really understands Chinese. So, according to functionalism, the Chinese room possesses real understanding.

Describe four replies to the Chinese Room argument (16 marks)

- **Impossibility Reply**: Criticise that the assumption that such a symbol manipulating program exists is not a nomic possibility, so the argument no longer holds since its assumption is false. ~~It is claimed that such a program to simulate the Chinese Turing test would likely be too long to run by hand or even digitally. This rejects the idea that the program is a nomic possibility, hence rejects the Chinese Room argument.~~
- **System's Reply**: Concede that the person/machine/components of the room do not individually understand Chinese, but they together make up a system that does. ~~Given that, the room as a whole understands Chinese. This rejects the assumption that a digital computer C will not differ from P in any sense that makes it more intelligent than P, as the computer C comprises a system that is intelligent, hence rejects the Chinese Room argument.~~
- **Virtual Mind (VM) Reply**: Concede that the person/machine/components of the room do not individually understand Chinese, but the whole system may create a virtual entity, distinct from the room and its parts, that understands Chinese. ~~This rejects the assumption that a digital computer C will not differ from P in any sense that makes it more intelligent than P, as the computer C is considered a separate entity in it of itself, hence rejects the Chinese Room argument.~~
- **Robot Reply**: Concede the person/machine does not understand Chinese, but if they were let out into the world, and were exposed to real application of Chinese, they could then learn it and understand its meaning by interacting with the world and others. ~~This rejects the assumption that a digital computer C will not differ from P in any sense that makes it more intelligent than P, as the computer can build on its existing knowledge and extend it, hence rejects the Chinese Room argument.~~

Four marks are awarded for just the definition, not including the extra ~~striken~~ text.

11. Essay Q: Ethics and Morality

Briefly summarise the main features of how Deontology, Virtue Ethics and Utilitarianism guide ethical behaviour (12 marks)

- Deontology says that we should follow a set of denotic rules (things that are permitted, obliged, prohibited) in order to behave ethically. There are two broad approaches; those that focus on the duties of agents to act, and those that focus on rights. In either case, the focus is on following a set of rules or laws that encode in them moral duties and respect for rights.
- Virtue Ethics says we should follow a set of virtues which then help us decide whether any given action is 'virtuous' and hence whether we should consider those actions if we want to behave ethically. The idea is that morally good actions flow from the cultivation of good character, which consists in the realisation of specific virtues.
- Utilitarianism is a special case of consequentialism that focuses on maximising the happiness of everyone involved, essentially we guide ethical behaviour by only considering to take actions that increase the 'global happiness' of the population affected.

Briefly describe two criticisms for each of Deontology, Virtue Ethics and Utilitarianism (12 marks)

Deontology:

- We can't possibly encode every single rule to cover every single case, it is physically impossible to cover all ethical behaviour using deontological rules.
- As a consequence of ↑: It's too general/vague and so doesn't guide action in specific situations.
- Deontic rules may prohibit things that aren't necessarily ethically negative, for example, we may be prohibited to speed on a road but it may be necessary if we are an ambulance trying to get to a hospital within a set time frame.
  As such it can be considered inflexible as it provides absolute prohibitions on certain actions and we may also consider that Deontologists usually fail to specify which principles should take priority when rights & duties conflict, thus Deontology cannot provide complete moral guidance.
- Applications of rules mean that overall consequences of actions may be harmful.
- Summary of topics:
  - Too general/vague and inflexible
  - Conflicts in duties and rights

Virtue Ethics:

- Different people, cultures, and societies have vastly different opinions on the consitution of a virtue, and as such there is no objectively right list.
- Does not provide guidance on what sorts of actions are morally permitted/disallowed, but rather qualities someone should cultivate to become a good person.
- Summary of topics:
  - Subjectivity of virtues
  - Lack of action guidance

Utilitarianism:

- Utilitarianism is **not demanding enough** and sometimes leads to intuitively unethical choices.
- Utilitarianism is **too demanding** and requires sacrifices that are unreasonable. Pragmatic utilitarianism says being a perfect utilitarian is anti-utilitarian since it is incompatible with the way our brains are designed.

Explain what is meant by ensuring accountability when AI algorithms are used to make decisions (6 marks)

Ensuring accountability means ensuring one can demonstrate to the public and to legal and regulatory bodies, that automated decisions comply with standards of legal and ethical fairness

OR

Ensuring accountability means ensuring that one can assign repsonsibility or blame when use of an algorithm results in an ehtically undesirable outcome (when things go wrong)

~~Accountability is the ability to place responsibility of decisions to an entity, whether a company, person, or otherwise. This means being able to understand why an AI makes certain decisions which is related to the problem of:~~

- ~~Understanding design decisions made by the people who made the AI and whether they were made in the interest of good ethical behaviour~~
- ~~Understanding why certain data was used to train the AI and whether it was chosen in the interest of good ethical behaviour and considerations were made about bias~~
- ~~Understanding how the AI itself reasons about certain ethical issues — however this is subject to the **black box problem** where we don't really understand how complex neural networks come to the conclusions that they do~~

~~Key aspects of ensuring accountability:~~

- ~~Transparency~~
- ~~Procedural regularity~~
- ~~Software verification~~
- ~~Cryptographic commitments~~
- ~~Adhering to ACM Principles~~

What does it mean for AI algorithms to be transparent, and why is transparency important for ensuring accountability? (10 marks)

For an algorithm to be transparent requires that one can examine and explain why an algorithm made any given decision. [4 marks] This is important for accountability as transparency allows one to demonstrate that the algorithm complied with ethical and legal standards [4 marks], AND transparency incentivises companies that design and use these algorithms to comply with the legal/ethical standard [3 marks],.

Transparency in AI algorithms means we can:

- Assess the data being used to train the AI and any biases in the data
- Assess the algorithm code/components and as such any design decisions made here
- Assess how the algorithm is reasoning and making the decisions it is
- Assess the contexts within which the AI is being used and whether the use is permitted or otherwise subject to any biases
- Hence, ensure algorithmic decisions are not biased/disciriminatory
- Incentivise ethical behaviour
- If we value procedural regularity, it would allow us to ensure the same policy/rule was used for every decision & that the decision-making is reproducible

Transparency is fundamentally required to understand whether the reasoning of the AI algorithm is faulty, biased, or otherwise sound, and hence drive decisions on whether the people building or using this AI algorithm are liable for any decisions that the AI makes.

A deontologist and utilitarian debate the importance of ensuring that algorithms are transparent. The deontologist claims that all algorithms should be fully transparent. What argument might the utilitarian use to challenge the deontologists' claim (in your example, give an example that the utilitarian could use to support their argument) (10 marks)

The utilitarian argues that full transparency of algorithms might not be ethically appropriate. Rather, one should weigh up the harms and benefits that would result from making any given algorithm fully transparent [4marks]. For an appropriate example [6 marks]. You might select one from the following:

1) Full transaprency might result in revealing private data about an individual and so violate GDPR regulations.

2) Full transparency of an algorithm that decides which tax returns to audit, might be exploited by a tax cheat to "game" the algorithm and avoid being audited.

3) Full transparency of an algorithm that selects who to pull aside for further checks at airport security might be exploited by a terrorist who, knowing the features of individuals that the algorithm identifies as suspicious, ensures that they do not display any of these features.

Making algorithms fully transparent poses some challenges:

- Data exposure: to prove transparency in a system, we may need to provide data being used as input into the AI algorithm which may include sensitive user information; as such, someone could abuse transparency of a system to exfiltrate sensitive information
- Potential gaming of systems: if an algorithm is transparent, it may be easier to cheat the system, for example if an AI is being used to assess whether fraud is occurring in a situation; as such, someone can abuse the transparency to learn how to not get detected as fraud

Both of these arguments can be used to challenge the deontologist, by stating that full transparency opens up vectors of abuse, with possible data exfiltration and other poor outcomes.