

LECTURER:
MARK GILLAN

MODEL ANSWERS:

:::AI ALIGNMENT

::: :::AI EXPLOSION

MODEL **ANSWERS** FOR EXERCISE: AI ALIGNMENT AND AI EXPLOSION

- Model answers for each of the questions
- Each question and answer on separate pages ...



QUESTION 1 – AI ALIGNMENT PROBLEM

- **Model answer:**

The AI alignment problem refers to the challenge of ensuring that artificial intelligence systems behave in ways that match human values and intentions. Even if an AI system follows its instructions correctly, it may still cause harm if its goals are not properly aligned with what humans want.

QUESTION 2 – RISKS OF THE AI ALIGNMENT PROBLEM

- **Acceptable points include:**

- AI prioritising efficiency over safety or fairness.
- Harm caused at large scale due to automation.
- Reduced human oversight as systems become more autonomous.

- **Model answer:**

One risk of the AI alignment problem is that an AI system may focus on efficiency and ignore human wellbeing, such as worker safety or fairness. Another risk is that small errors in the goals given to AI can be repeated at scale, causing widespread harm before humans can intervene.

QUESTION 3 – OPPORTUNITIES OF AI ALIGNMENT

- **Acceptable points (at least two explained) include:**

- Safer and more ethical AI systems.
- Increased public trust.
- Better use of AI to solve complex problems.

- **Model answer:**

Solving the AI alignment problem can lead to safer and more ethical AI systems that support human decision-making. It can also increase public trust in AI, allowing the technology to be used responsibly to solve complex problems such as healthcare or climate change.

QUESTION 4 – AI EXPLOSION

- Definition of an AI explosion.
- Reference to rapid or accelerating AI development.

- **Model answer:**

A potential AI explosion refers to a situation where artificial intelligence develops very rapidly, possibly improving itself faster than humans can control or regulate. This rapid growth could lead to sudden and significant changes across society.

QUESTION 5 – RISKS OF A POTENTIAL AI EXPLOSION

- **Acceptable points (at least 2 should have been done) include:**

- Loss of human control.
- Inability of laws and regulation to keep pace.
- Sudden economic or social disruption.

- **Model answer:**

One risk of a potential AI explosion is that humans may lose control over AI systems as they develop faster than regulation can keep up. Another risk is sudden economic disruption, where jobs and industries change too quickly for workers to adapt.

QUESTION 6 – OPPORTUNITIES OF A POTENTIAL AI EXPLOSION

- **Acceptable points (at least 2 should have been done) include:**

- Faster scientific discovery.
- Medical breakthroughs.
- Increased productivity and efficiency.

- **Model answer:**

An AI explosion could lead to rapid scientific and medical breakthroughs, such as faster drug discovery. It could also increase productivity across industries, improving living standards if managed responsibly.

QUESTION 7 – CONNECTION BETWEEN ALIGNMENT AND EXPLOSION

- Explanation of why alignment becomes more important.
- Clear link between speed and risk.

- **Model answer:**

AI alignment becomes more important during an AI explosion because rapid development increases the risk of harm if systems are not properly controlled. If AI grows quickly without being aligned to human values, mistakes or harmful behaviour could spread faster and become harder to stop.



THE END

Research, check knowledge
again, text yourself