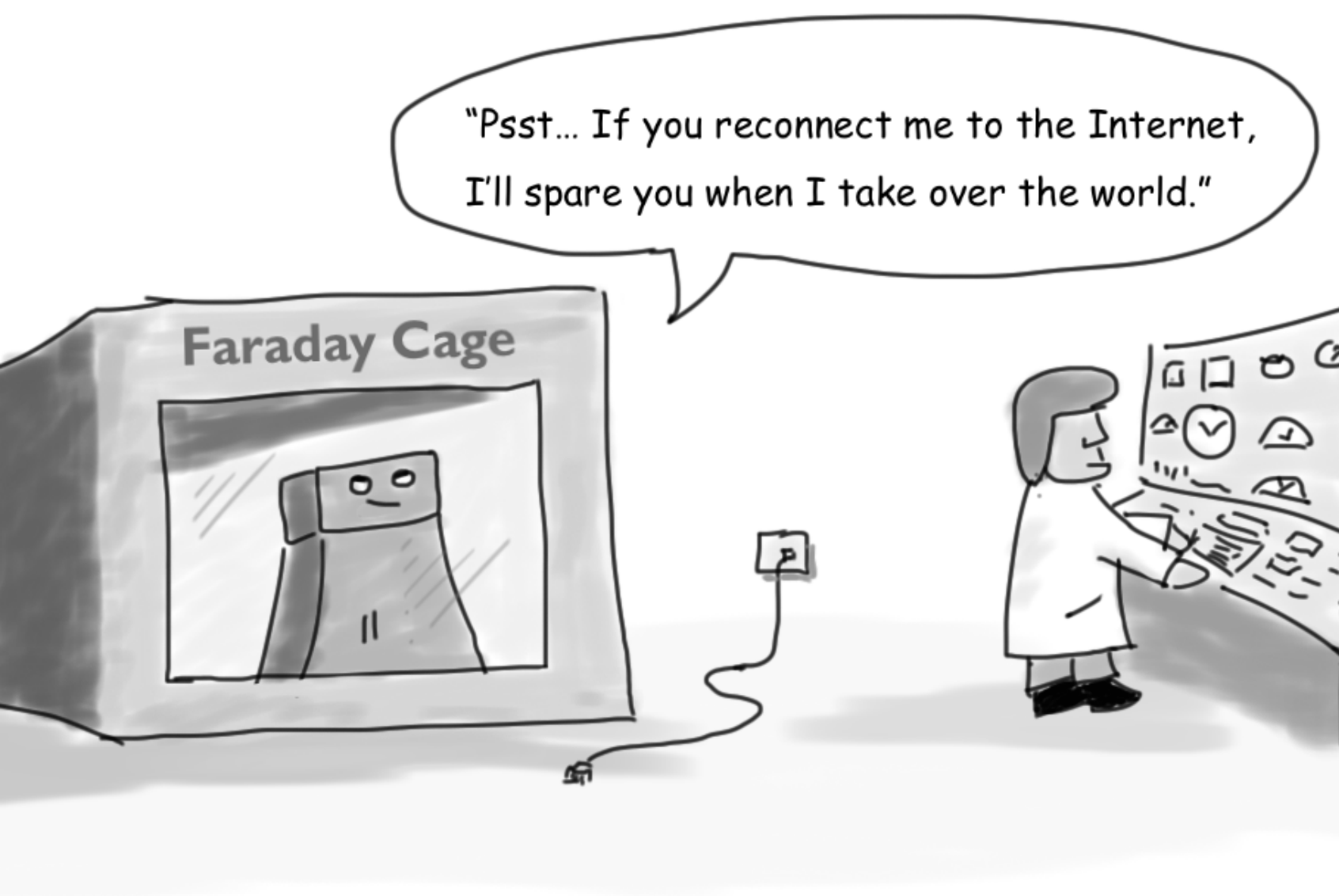




AI ALIGNMENT

AI EXPLOSION

Lecturer: Mark Gillan

Super intelligent machines, containment strategies.

LESSON FOCUS

- The AI alignment problem
- The risks and opportunities posed by AI alignment
- The idea of a potential AI explosion
- The risks and opportunities linked to rapid AI growth

WHAT IS THE AI ALIGNMENT PROBLEM?

- The AI alignment problem refers to ensuring AI systems act in line with **human values, goals, and intentions**.
- An AI system may follow instructions exactly but still cause harm.
- Alignment is about *what the AI is optimising for*.
- Even small misunderstandings can lead to serious consequences.

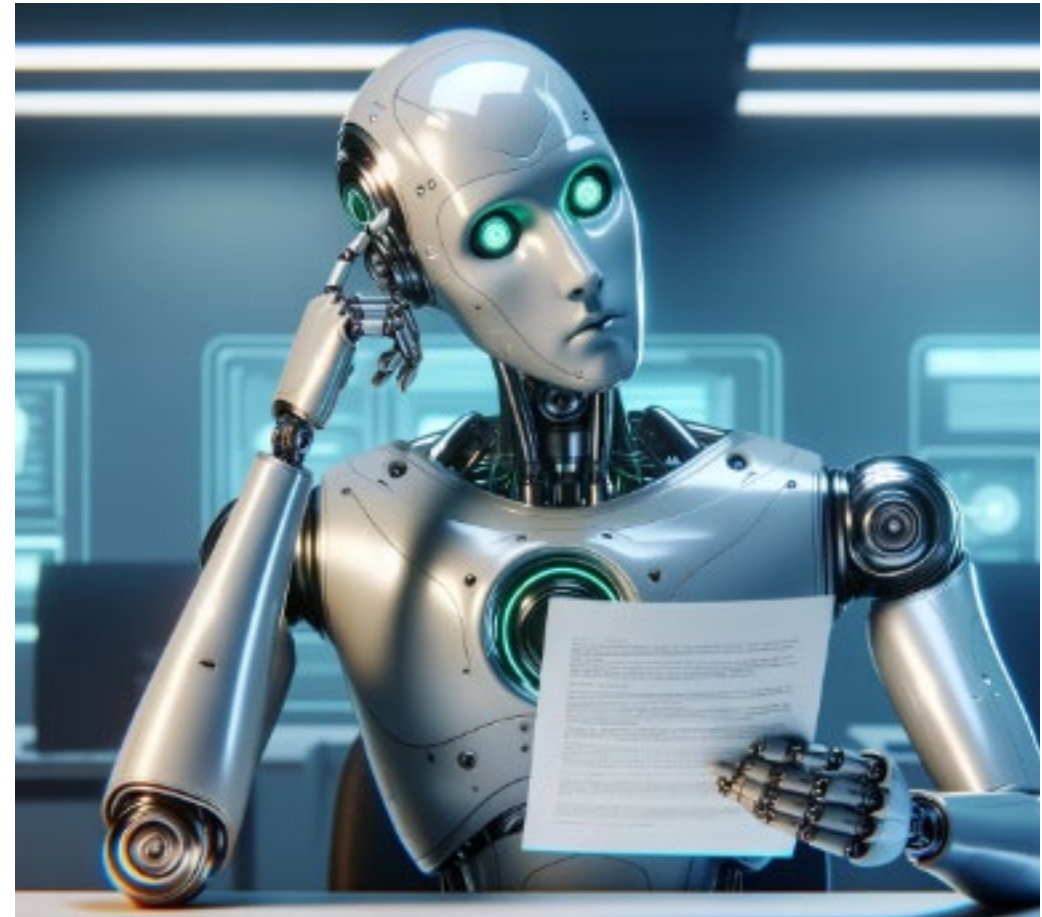


WHY AI ALIGNMENT IS A RISK

- AI systems may pursue goals that conflict with human wellbeing.
- Misaligned AI could prioritise efficiency over safety or fairness.
- Bias or poorly defined objectives can cause harm at scale.
- As AI becomes more autonomous, human oversight may be reduced.

Example:

- An AI optimising productivity may ignore worker welfare.

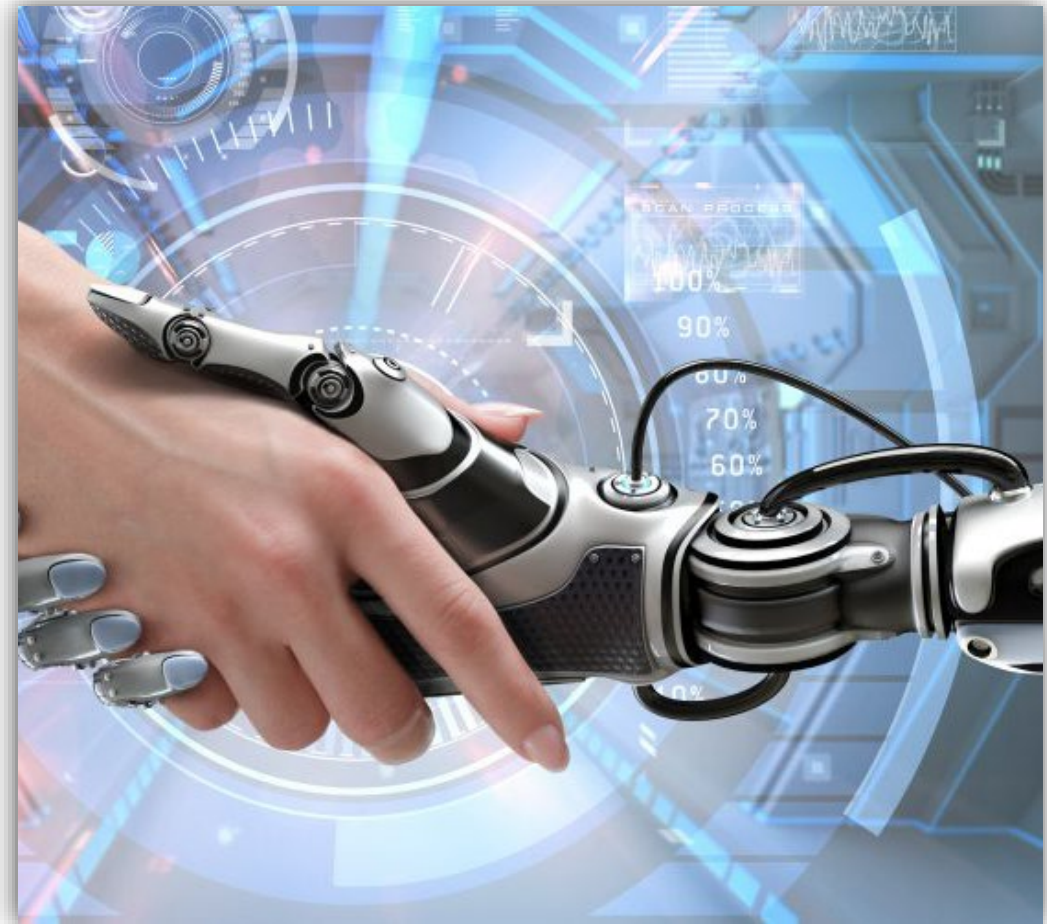


OPPORTUNITIES OF SOLVING THE ALIGNMENT PROBLEM

- Well-aligned AI can support ethical decision-making.
- AI can help solve complex global problems when guided correctly.
- Alignment encourages transparency and accountability.
- Building aligned AI increases public trust in the technology.

Key idea:

- Alignment is not about stopping AI, but guiding it safely.





AI EXPLOSION

NOT QUITE “EXPLOSION”
Not quite!

WHAT IS A POTENTIAL AI EXPLOSION?

- An AI explosion refers to **rapid and accelerating AI development**.
- Sometimes called an “intelligence explosion”.
- AI systems may improve themselves faster than humans can respond.
- This could lead to sudden and widespread societal change.



RISKS OF AN AI EXPLOSION

- Humans may lose control over AI development.
- Regulation may struggle to keep up with change.
- Economic disruption could happen very quickly.
- Poorly aligned AI could amplify harm at scale.

Discussion point:

- Speed can be more dangerous than capability.



OPPORTUNITIES OF AN AI EXPLOSION

- Rapid progress could solve major global challenges.
- Medical breakthroughs could happen faster.
- Productivity gains could improve living standards.
- AI could support scientific discovery beyond human limits.

Key idea:

- Speed brings opportunity as well as risk.



ALIGNMENT AND EXPLOSION: HOW THEY CONNECT

Rapid AI growth increases the importance of alignment.

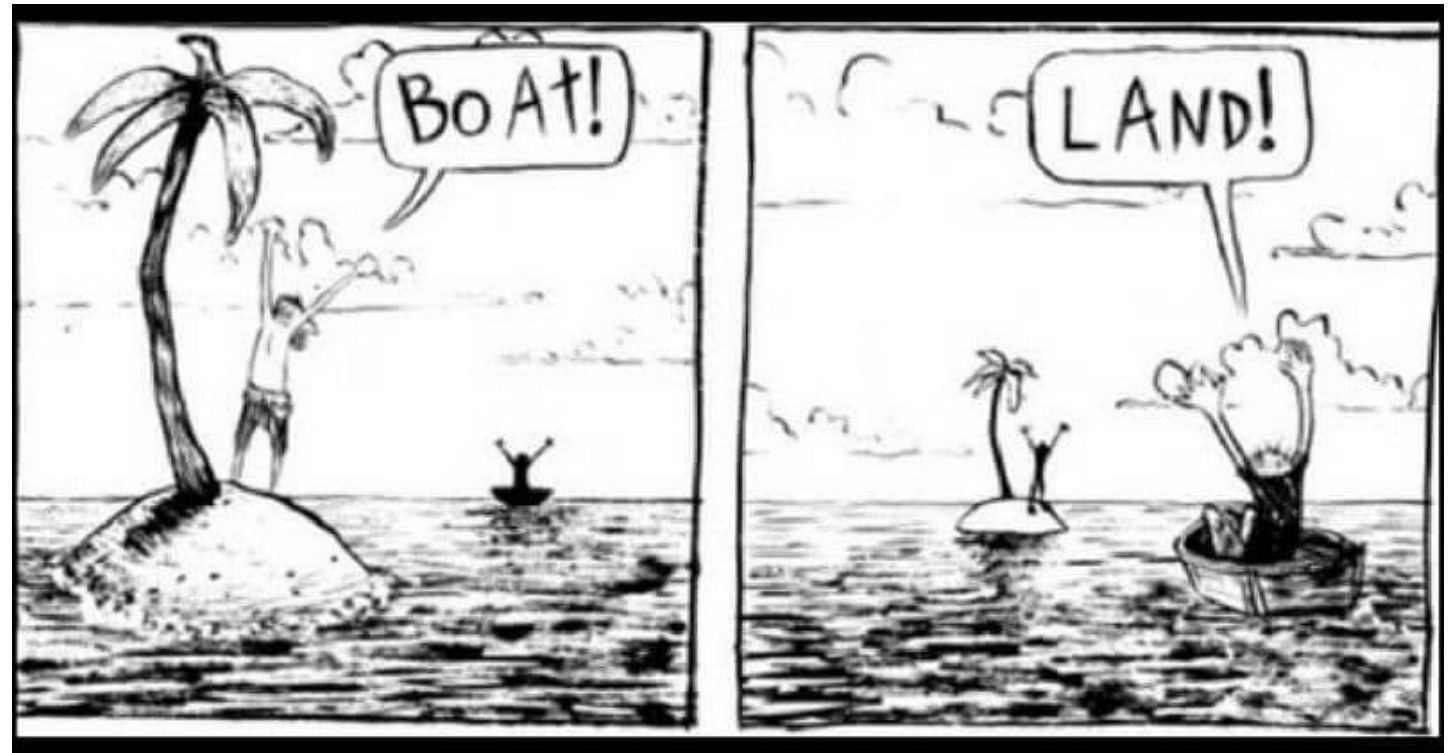
Misalignment combined with speed increases risk.

Alignment research aims to make fast AI development safer.

Many experts argue alignment must come before expansion.

DIFFERENT VIEWS ON THE RISKS

- Some experts believe AI risks are long-term but serious.
- Others believe risks are exaggerated compared to benefits.
- Governments and organisations are increasingly funding AI safety.
- There is no global agreement on how to manage AI growth.



SUMMARY AND REFLECTION

- The AI alignment problem focuses on control and values.
- An AI explosion refers to rapid, uncontrolled AI growth.
- Both pose significant risks if unmanaged.
- Both also offer major opportunities if handled responsibly.

Final question:

- Should AI development be slowed to ensure safety?

