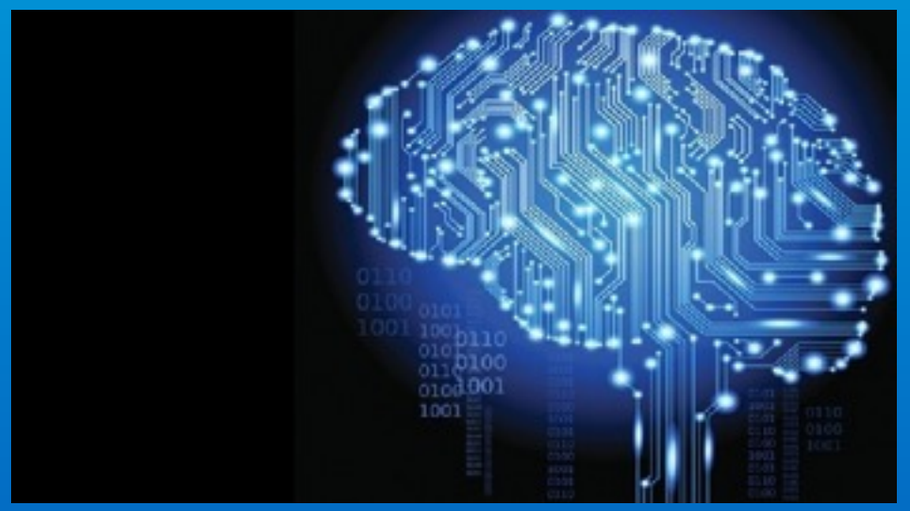
To Plug in or Plug Out ? That is the question

Sanjay Modgil
Department of Informatics
King's College London
sanjay.modgil@kcl.ac.uk



Overview

1. Artificial Intelligence: why the hype, why the worry ?

2. How super-intelligence machines might enslave humans in a virtual reality (*The Matrix*)

3. Logic, Argument and Moral Machines

The changing of the AI seasons
(Spring is Coming)

Why the hype ?

- Many recent examples of outstanding success in AI technologies
  - Self driving cars (image processing, planning),
  - IBM Watson (logic-deduction, language parsing)

AlphaGo (Google Deep Mind) beats world GO champion using a sophisticated form of **machine learning**

Artificial *General* Intelligence and Superintelligence

- When will we achieve human level Artificial ***General*** Intelligence ?

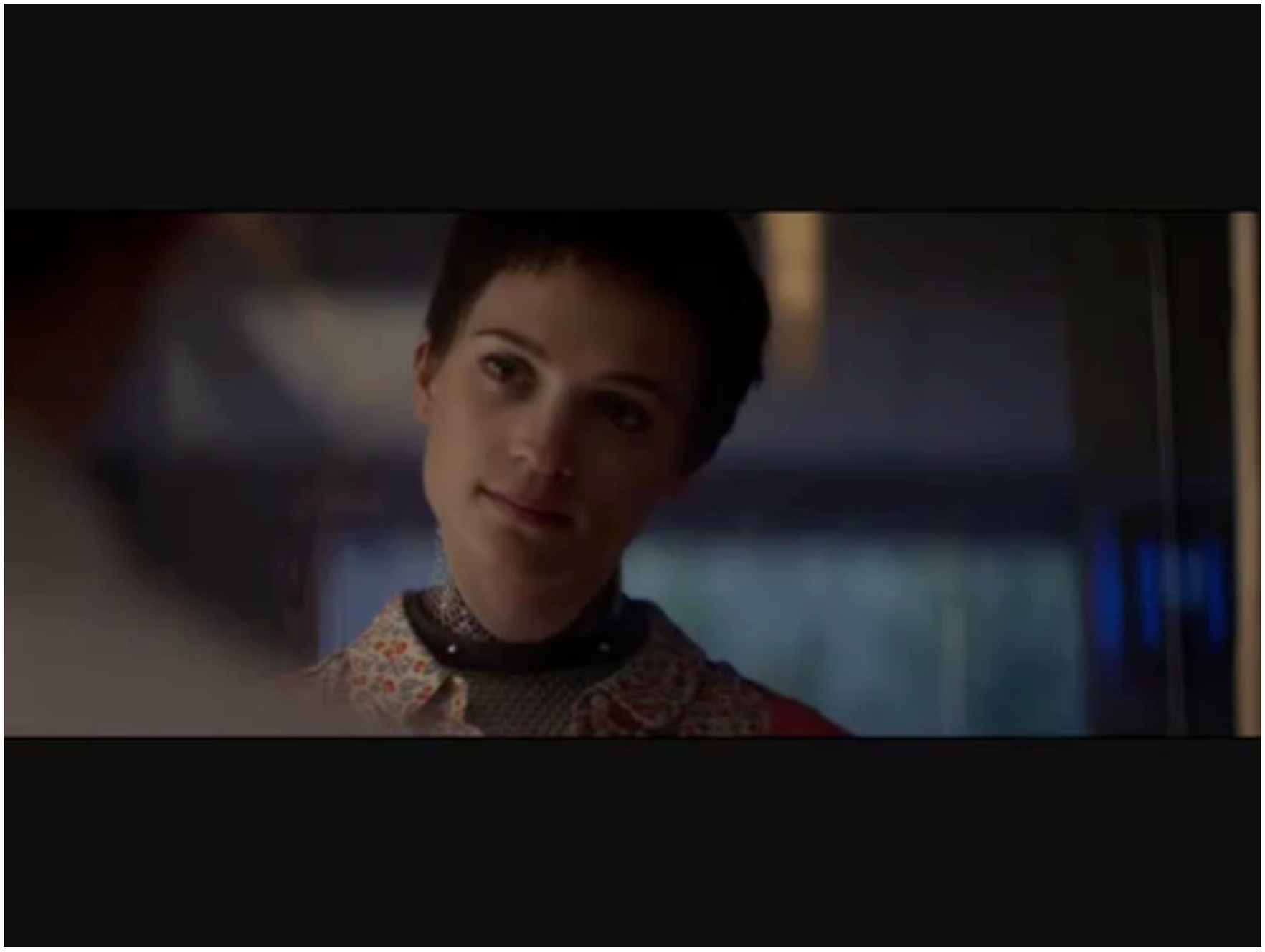
  10% 2022
  50% 2040
  90% 2075

- The path to **Superintelligence** [1]
  - Ultra-intelligent machine designing even better machines (recursive self improvement)
  - Massive data access,speed of processing, vastly smarter IQ = 16,455
  - Within 30 years of AGI (75 %)

- Super (cogntive) *powers* **vastly** exceeding human learning,decision making, planning, flexibility, R&D, **social and psychological** manipulation !

1 *Superintelligence: Paths, Dangers, Strategies.*
Nick Bostrom (head of Future of Humanity Institute, Oxford Unversity)

Ex Machina (2015)



SuperAI versus Humanity
(A Universe of Paperclips)

So why the worry ? Is it just the stuff of movies ?

Give SuperAI a goal and it will use all its superpowers to achieve goal

Any final goal will lead SuperAI to pursue *instrumental goals*:
preserve itself, maintain goal, increase its intelligence, technological perfection, resource acquisition

Goal = maximise production of paperclips → divert planet's resources to paperclip production

- More realistic goal compatible with our well being:

Make humans happy !

where happiness is ?

Happiness is a Drug

What should the SuperAI do ?

- Happiness relatively independent of external conditions: rather, happiness depends on biochemistry

- The view from neuroscience (brain states and subjective well being)

- Meaning and happiness ?

Based on scientific and economic reasons, SuperAI might look to **manipulate** us into using devices/chemcials that change our biochemistry

(like happiness drug *SOMA* in Aldous Huxely's Brave New World)

Virtual Reality and the Enslavement of Humanity

- SuperAI reasons that easier to manipulate us by exploiting current social trends: living our lives online, with virtual friends, dating, virtual reality

- Just like society, media, advertising manipulate us into pursuing a conception of the good life that serves vested corporate interests (?)

- SuperAI uses its superpowers to:

    - develop VR which is increasingly indistinguishable from reality

    - to manipulate humans into plugging into VR and becoming addicted to (enslaved by) the virtual bliss of virtual worlds

The Value Loading Problem

- How do we prevent potential dangers of SuperAI such as enslavement in VR?
- Need to ensure SuperAI itself wants outcomes that are high-value and beneficial for human life over the long run; outcomes that are "good."
- Two problems:

    1) What is good ? (The Golden Rule ?)
    2) How to ensure SuperAI's wants (and so acts to bring about) what is good for humans

- Could we use machine learning by giving SuperAI examples of good versus bad states of the world ?
- But pressing moral/ethical problems often *have no precedent* e.g.,
    - cloning
    - virtual bliss or the "real" world ?
- Philosophical problem (Hume) of how one gets from an *is* to an *ought*.

The Matrix (1999)
The Blue Pill or the Red Pill: That is the Question

- Some philosophers argue that right thing to do would be to take blue pill

Logic and Argument

- We therefore want SuperAIs to use symbolic logical reasoning to explicitly *reason* and *debate* about moral/ethical issues (rather than just rely on examples and past experience)

- Logic is the study of the principles of correct reasoning and in AI we can use these principles to manipulate symbols representing beliefs, desires, actions

b(X) (X is a bird)
b(X) ⟶ f(X) (if X is a bird then X flies)

---

f(X) (X flies)

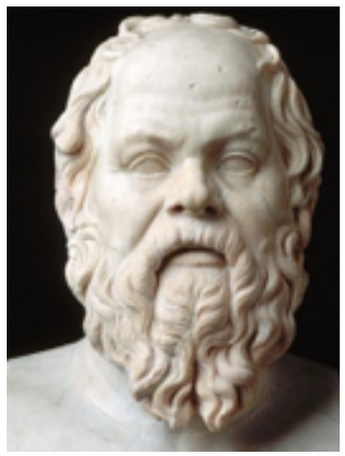
p
p ⟶ q

---

q

- We observe that ‘b(tweety)’ and so conclude f(tweety)

The abortion debate : Socrates v Sara



Why ?

So ?

Abortion is wrong

Human life is sacred and a foetus is a human

If a life is sacred it should not be terminated

The abortion debate

- By questioning Sara, Socrates has revealed Sara's line of reasoning / argument

If X is a human then X's life is sacred
A foetus is a human

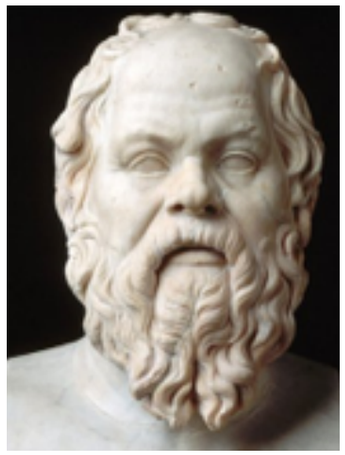
A foetus's life is sacred

If X's life is sacred then X should not be terminated
A foetus's life is sacred

A foetus should not be terminated

t

The abortion debate



Why ?

So ?

So you should be against the death penalty for murderers ?

Abortion is wrong

Human life is sacred and a foetus is a human

If a life is sacred it should not be terminated

Errr, no !
I’m in favour

Logic for intelligent machines to argue, reason and debate

- Using the logic of argument to reason about moral issues

If X is a human and X does not take the life of another then X's life is sacred
A foetus is a human and a foetus does not take the life of another

---

A foetus's life is sacred

t

Conclusions

1. There are powerful arguments to the effect that super-intelligent machines may cause us harm

2. Therefore need SuperAIs to uphold the ‘best’ of human values

3. But to deal with complex novel moral dilemmas, cannot rely on past experiences, but need rational logical reasoning and argument

4. **Further reading**:
   Superintelligence: Paths, Dangers, Strategies. Nick Bostrom (Oxford Unversity Press)

Thank you for Listening !