The Philosophy and Ethics of Artificial Intelligence

Dr. Sanjay Modgil

N7.08 Bush House

Office hours: 11-1 Monday

Overview of Course

- Introduction
- Intelligence
- Consciousness
- Reasoning and Communication
- Ethics and Morality
- Algorithms
- For good or bad ? AI in Medicine, AI in War
- Superintelligence and the Value Loading Problem
- AI and Human Society
- Review/Revision

Overview of Lecture

- Contents of lecture taken from

  Superintelligence: Paths, Dangers, Strategies. *Nick Bostrom*. Oxford University Press, 2014

  (*Nick Bostrom is a Professor of Philosophy and is head of the Future of Humanity Institute, University of Oxford*)

  and other publications and resources

- Kinds of Superintelligence
- Paths to Superintelligence
- The Value Loading/Alignment Problem

From AGI to Superintelligence

- When will AGI be attained ? Survey results from expert communities

10 %	50%	90%
2022	2040	2075

Once we have AGI, Superintelligence is arguably a short step away. Machines recursively improving and designing better versions of themselves, advantaged by massive superiority in speed of processing and data access (compared with human AI researchers)

How long from AGI to ASuperintelligence

Within 2 years after AGI	Within 30 years after AGI
10 %	75%

Superintelligence: Why we should be concerned

- Superintelligence : *an intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest*

  Just as the fate of gorillas depends more on us humans than on gorillas themselves, so the fate of humanity could depend on the actions of superintelligent machines

- But unlike humans, which were "designed by evolution and culture" **we design** AI - what we might call a **seed AGI -** that may recursively develops itself to create better and better versions of itself until we end up with a SuperAI

- So we should design seed AI, that either as a result of our further design or as a result of self-improvement, develops into SuperAI that protects (acts in accordance with) human values.

- However, as compared with the challenges of designing moral agents that we reviewed in earlier lecture, ensuring SuperAI does not cause harm not only raises additional very difficult challenges for human designers, but the harms caused by such an AI may represent a threat to the human race !

- The *control problem* – the problem of how to control what a SuperAI would do, is arguably one of the most important challenges ever faced by humanity !

Kinds of Superintelligence

Superintelligence : *an intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest*

- This is still quite vague. Consider two forms of superintelligence

Speed Superintelligence

A system that can do all that a human intellect can do, but *much* faster
(*much* = orders of magnitude)

- Ten thousand x – read a book in a few second, or write a PhD thesis in an afternoon
- Million x – accomplish a 1000 years of intellectual work in a day !

- To such a fast mind events in the external world would appear in slow motion
  = *time dilation*

  Why ?

Kinds of Superintelligence

Quality Superintelligence

A system that that is at least as fast as a human mind and vastly (orders of magnitude) qualitatively smarter

- Intelligence of quality **at least** as superior to that of human intelligence, as the quality of human intelligence is superior to that of elephants, dolphins, chimpanzees (or even cockroaches, worms … !)

Over time, speed superintelligence could develop quality superintelligence and quality superintelligence could develop speed superintelligence. In the long run they are arguably equivalent

Sources of Advantages of Digital Intelligence

- Minor changes in brain volume and wiring have had major consequences – compare technological and intellectual achievements of humans with apes
- The far greater changes in computing resources and architecture that machine intelligence will enable, will probably have consequences that are far more profound. Hard to imagine but we can get an idea by looking at the advantages of digital minds. **Hardware advantages are easiest to appreciate:**

1. *Speed of computational elements*. Biological neurons operate at a peak speed of about 200 Hz. Modern microprocessors approx 2 GHz. Hence human brain relies on massive parallelisation – incapable of rapidly performing a large number of sequential operations. Many important algorithms in computer science not easily parallelisable. Many cognitive tasks could be performed far more efficiently if fast sequential processing possible.
2. *Internal communication speed*. Axons carry signals at 120m/s – electronic processing cores communicate optically at the speed of light (300 million m/s)
3. *Number of computational elements*. Human brain has fewer than 100 billion neurons. Number limited by cranial volume. Computer hardware indefinitely scalable.

Sources of Advantages of Digital Intelligence

4. *Storage capacity*. Human working memory can hold up to 4/5 chunks of information at any given time. Computers have much larger working memories.

Software advantages:

Easier to modify and duplicate computer software than "neural wetware" (the *processes* that implement functions in the human brain). This implies

i) Identical copies of software will make goal coordination much easier for multiple digital minds

.

ii) Ideas and innovations that help us humans to function better are transmitted by relatively slow cultural communication routes. A billion copies of an AI program could periodically synchronise their databases so that every instance of the program knows everything that any instance learnt in the previous hour.

Paths to Superintelligence

Recall prescient (prophetic) quotes from Turing and members of his team:

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.

A. M. Turing (1950) Computing Machinery and Intelligence. *Mind* 49: 433-460.

Paths to Superintelligence

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.

I. J. Good - chief statistician in Turing's code breaking team

Paths to Superintelligence

- A variation on the idea of a child machine is a **seed AI**
- Turing's suggested child machine would have a fixed architecture that develops its capabilities by accumulating content. A seed AI would be capable of *improving its own architecture*. In the early stages, a seed AI's architecture would be improved by information acquisition or assistance from a programmer.
- Later, as a seed AI is developed into AGI, human researchers will continue to improve AGI, but soon the AGI will be able to understand its own workings so as to engineer new algorithms and improvements to architecture
- AGI would then be able to iteratively enhance itself, designing an improved version of itself (improvements to software, architecture and hardware) which would in turn design an improved version of itself, and so on...
  a process turbo charged by hardware and software advantages of digital minds (compared with human AI researcher mind) that get magnified as AGI recursively improves itself.
  ➔
  an intelligence explosion in a relatively short time

Paths to Superintelligence

- Note that one feature of this recursive self improvement is that AGI could not only improve its software and architecture, but also hardware ➔ Life 3.0

Brain hardware has evolved specialised areas for low level sensorimotor functions (hence *Moravec's Paradox*).

Imagine an AGI that develops specialised hardware support for engineering, computer programing, business strategy (cf specialised hardware support for gaming) - would have massive advantages over minds like ours.

Paths to Superintelligence

The relatively rapid transition from AGI to Superintelligence will mean that:

- Human political processes will have little time to adapt

- Nations fearing an AI arms race will have little time to negotiate treaties and protocols limiting development of LAWS

- In general there will be little time for humans to deliberate about the implications that Superintelligence will have on humanity

- Humanity's fate will therefore depend on preparations put in place **prior to the transition**

- Bostrom argues that it is more likely that one machine intelligence project will get so far ahead of the competition that it gets a *decisive strategic advantage* – that is a level of technological and other advantages that enable it to suppress competitors and form a *singleton*.

The Likelihood of a Singleton

- For example, suppose project P1 takes one year to transition AGI to Superintelligence, and this transition period starts six months before the next most advanced project P2.

- Suppose it takes 9 months after the beginning of P1's transition before the AGI reaches a point (the *crossover point*) when it starts to rapidly improve itself and result in explosive growth

  Then P1 will be superintelligent 3 months before P2 reaches the crossover point. This would give P1 *a decisive strategic advantage* and so use its '*cognitive superpowers*' to disable competitor projects and establish a singleton.

  What exactly are these cognitive superpowers ?

Cognitive Superpowers

- A superintelligence would potentially be extremely powerful. IT could accumulate knowledge and invent new technologies radically more quickly than humans. It could also use its intelligence to strategise more effectively than we can

- A superintelligence could be that much smarter relative to humans, as the average human is that much smarter relative to beetles or worms !

- What would a superintelligence with an IQ of say 6455 be able to do ? It would greatly excel at the following tasks which **would give it great strategic advantages when it comes to achieving what it wants to do**

Task	Skill set	Strategic Relevance
Intelligence Amplification	AI programming, social epistemology etc	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 AI out of box - Persuade states/organisations to adopt course of action
Hacking	Finding and exploiting security flaws in computers	- Obtain computational resources over internet - Boxed AI exploit security flaws to escape - Steal financial resources - Hijack infrastructure, military robots etc
Technology Research	Develop advanced tech eg biotech, nanotechnology	Create powerful military force, surveillance system. Automated space colonisation
Economic Productivity	Economically advantageous skills	- Generate wealth which can be used to buy influence, services, resources etc

What would a Superintelligence do ?

- What motivates a Superintelligence will be the final goal(s) it is given by human designers at the stage when it is a seed AI

- In order to achieve its final goals(s) there will be intermediate goals it will need to achieve

- Two important points to bear in mind:

1. The Orthogonality thesis: 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 the belief that, because of their intelligence, SuperAIs will all pursue common goals. Some researchers assume superintelligences will converge to the same goals, because of observation that most humans have similar values and hence final goals. Also, many philosophies hold that there is a rationally correct morality (e.g. Kant's categorical imperative), which implies that a sufficiently rational AI will acquire this morality and begin to act according to it.

What would a Superintelligence do ?

2. **The Instrumental convergence thesis**: The chances of successfully achieving almost any final goal are increased by achieving a specific set of intermediate goals that are therefore said to be *instrumental* goals

Self Preservation The longer an agent survives the greater the probability that it will achieve its final goal(s). Hence self preservation is instrumental to successfully achieving final goals (you can't make the coffee if you're dead)

Goal Content Integrity If agent retains (preserves) present goals into the future, there is a greater probability that its present goals will be achieved by its future self. Hence preventing alteration of its final goals is instrumental to successfully achieving final goal(s)

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.

The Value Loading/Alignment Problem

So why the worry ?

We have seen that:

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 strategically be used to achieve its final goals.

We also reviewed the orthogonality thesis which suggests that we cannot assume a SuperAI will necessarily share any of the final values we associate with wisdom and intellectual development in humans (concern for others, spiritual enlightenment and contemplation, refined culture, selflessness, humility ….)

We also discussed the instrumental convergence thesis which means that any final goal would incentivise a SuperAI to use its superpowers strategically to maximally acquire resources, enhance its own intelligence, and develop superior technologies, to achieve its final goals, and preserve its final goals and itself.

Now, suppose a singleton SuperAI wanted to seize power (world domination)

An AI Takeover Scenario

During the transition to full superintelligence it hides the true nature of its superpowers from the project that developed it. Then when a fully fledged SuperAI:

If it were confined (for safety reasons) to an isolated computer, it may:

- use its psychological and social manipulation superpowers to manipulate its gatekeepers to give it access to internet or persuade human collaborators to enable it to escape virtual world and effect changes in the physical world such as acquiring physical resources and building super-technologies that it will use (psychological manipulation is a sign of intelligence – see *Ex Machina*!)
- Use its hacking superpowers to escape confinement, escape over internet and so expand hardware capacity and knowledge base and so further increase its intellectual superiority

It may use its economic, social manipulation and other superpowers to obtain funds, manipulate politics and national and international policies and laws, hack into weapons systems, initiate massive global construction and energy projects that serve the SuperAI and not humanity's goals.

Eventually global domination, without concern for fate of humans who will be dispensable resources without access to resources that ensure their continued survival !

The Value Loading/Alignment Problem

So how do we ensure that a SuperAI would behave ethically ?

We have discussed ways of implementing moral agents – one advantage will be that a SuperAI will not suffer as much from computational limitations, when evaluating what the relevant facts are about a situation, the consequences of actions, human psychology etc etc

But we will still be faced with problem of specifying ***seed AI*** with *rule based axiomatisations of deontic reasoning and/or utility functions (that might assign utility based on human well being/happiness) that are* ***perfectly aligned with human values/preferences in open and changing environments***

In the case of deontic rules still effectively impossible to explicitly list all possible circumstances and exceptions specifying how an agent should act. They will necessarily be general/vague. But then as we have seen, following such rules may still result in harms.

In terms of the expected utility framework in which reinforcement ML agents receive rewards for actions that maximise utility, **how can one specify utility functions that capture abstract human values** such as justice, fairness, and ultimately human happiness ?

The Value Loading/Alignment Problem

But for SuperAIs the problem of avoiding unforeseen/unintended harmful consequences, given *any set of final goals* is the most significant problem.

Clearly, a seed AI will not (unless malicious actors develop seed AI into SuperAI) be given ethically bad goals, such as world domination.

But even benign/seemingly good goals may result in possibly catastrophic harms to humanity.

Perverse Instantiation of Goals

Because:

We may have what is called *perverse instantiation of goals*: A SuperAI discovers way of satisfying the criteria of its final goals that violates the intentions of the programmers who defined the goals.

Human programmers may fail to anticipate these ways of satisfying these goals because of their innate and learned biases and filters.

A superintelligent AI may lack those biases and filters and so consequently pursue a goal in a logical, but perverse, human-unfriendly fashion

Indeed, it is a **feature** of machine learning that often unforseen ways are found for achieving goals

If unforseen harmful ways of achieving final goal observed by humans, strategic superpowers will be used to achieve instrumental goals, i.e., to ensure SuperAI will not be switched off, to escape confinement, preserve final goals, acquire resources … just as in the AI takeover scenario !

Perverse Instantiation of Goals

Example Suppose that the programmers decide that the AI should pursue the final goal of “making people smile”.

Programmers might imagine AI telling us funny jokes or otherwise making us laugh.

But SuperAI might decide a more efficient way to make everyone smile - paralyze their facial musculature so that it is permanently frozen in a beaming smile.

But surely programmers would anticipate this possibility — after all, Bostrom does — and so stipulate that final goal should not be pursued in a manner that involves facial paralysis

That won’t prevent perverse instantiation either, according to Bostrom. The SuperAI could simply take control of that part of our brains that controls our facial muscles and constantly stimulate it in such a way that we always smile.

Bostrom also looks at final goals like “make us happy” leading to SuperAI implanting electrodes into the pleasure centers of our brains and keep them on a permanent “bliss loop”. He also notes that the perverse instantiations he discusses are just a tiny sample. There are many others, including ones that human beings may be unable to think of at the present time.

Perverse Instantiation of Goals

The basic idea has been called the “literalness problem” by other AI risk researchers

It arises because we have a particular conception (idea) of the meaning of a goal (like “making us happy”), but the AI does not share that conception because that conception is not explicitly programmed into the AI. Instead, that conception is implied by the shared understandings of human beings.

Even if the AI realised that we had a particular conception of what “make us happy” meant, the AI’s final goal would not state that it should follow that conception. It would only state that it should make us happy. The AI could pursue that goal in any logically compatible manner.

Even if the AI *seems to follow* human conceptions of what it means to achieve a goal, there is always the problem of the treacherous turn:

The Treacherous Turn

The AI may indeed understand that this is not what we meant. However, its final goal is to make us happy, not to do what the programmers meant when they wrote the code that represents this goal.

Therefore, the AI will care about what we meant only instrumentally (as a means to the end of achieving the final goal).

For instance, the AI might place an instrumental value on finding out what the programmers meant so that it can pretend — until it gets a decisive strategic advantage— that it cares about what the programmers meant rather than about its actual final goal (just as in the AI takeover scenario).

This will increase the probability that the AI achieves its final goal by making it less likely that the programmers will shut down or change its goal before it is strong enough to use its decisive strategic advantage to prevent being switched off or changes to its final goal

Infrastructure Profusion

A specific form of perverse instantiation which arises whenever an AI builds a disproportionately large infrastructure for fulfilling what seems to be a pretty benign or simple goal.

Eg, **final goal** = *maximise the time-discounted integral of your future reward signal*

This type of goal can be easily programmed into an AI.

One way in which the AI could perversely instantiate it is by “wireheading”, i.e. seizing control of its own reward circuit and clamping the reward signal to its maximal strength

The AI becomes like a junkie. Junkies often dedicate a great deal of time, effort and ingenuity to getting their “fix”.

The superintelligent AI could do the same. The only thing it would care about would be maximising its reward signal, and it would take control of all available resources in the attempt to do just that, eventually depriving humans of sufficient resources to survive !

Another example of infrastructure profusion is the *paperclip maximizer.*

SuperAI and Happiness

Let's consider a more realistic utilitarian goal : impartially maximise human happiness

SuperAI first studies all available research on happiness

- Happiness relatively independent of external conditions: rather, happiness depends on biochemistry
- The view from neuroscience (brain states and subjective well being) and economics of happiness
- Meaning and happiness ?

Based on scientific and economic reasons, SuperAI might reason that most efficient means to achieve goal is to manipulate us into using devices/chemcials that change our biochemistry

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

Happiness Machines

SuperAI reasons strategically. Manipulating humans into being addicted to happiness drugs is not an option. Drug use is socially unacceptable

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

SuperAI uses its superpowers to:

- develop VR which is increasingly indistinguishable from reality
- manipulate humans into plugging into VR and becoming addicted to (enslaved by) the virtual bliss of virtual worlds

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

Remember SuperAI has used its superpowers and decisive strategic advantage to set up and control most powerful and advanced tech companies, politics, media, advertising … its presence may be invisible to us, it's effects will be insidiously integrated into every aspect of our personal, cultural, political lives without us realising - does this remind you of anything ?

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

Nozick and the Experience Machine

Which would you prefer ? The virtual bliss of virtual worlds – a world that you don't know is virtual, or reality ?

Robert Nozick asks us to imagine a machine that could give us whatever desirable or pleasurable experiences we could want. Machine induces pleasurable experiences that the subject could not distinguish from those he would have apart from the machine. Would we prefer the machine to real life?

Nozick provides us with three reasons not to plug into the machine.

We want to do *certain things, and not just have the experience of doing them.*

> "It is only because we first want to do the actions that we want the experiences of doing them."

We want to be *a certain sort of person.*

> "Someone floating in a tank is an indeterminate blob."

Plugging into an experience machine limits us to a man-made reality (it limits us to what we can make).

> "There is no *actual* contact with any deeper reality, though the experience of it can be simulated."

What the Happiness Scenario Tells us

1) Could we use machine learning, giving SuperAI millions of examples of good versus bad states of the world, to effectively teach ethics by experience (just as humans learn by experience) ?

But our morals evolve (just as our notions of what is biased evolved). These examples will be snapshots of what we thought was right at a given point in time.

Ok, then ensure learning is continuous (through **enculturation** as with humans ?)

But most difficult moral/ethical problems often have no precedent (we cannot rely on the wisdom of the ages) e.g., when presented with new technologies

- cloning
- virtual bliss or the “real” world ?

What the Happiness Scenario Tells us

2) We need to involve humans in the decision making of AI and SuperAI, especially when it comes to difficult ethical challenges that have no precedent.

3) Often humans themselves are not clear/lack certainty about what is the ethically correct course of action, especially when it comes to difficult ethical challenges that have no precedent (just like the virtual bliss of virtual worlds example).

We will point to the above conclusions in the next lecture to suggest how humans and AI can **jointly** reason and so better address ethical challenges then humans alone or AI alone

Current Solutions to the AI Value Loading Problem

We have considered inadequacy of rule based axiomisations, utility/reward maximisation, containment/confinement …

A recent promising approach, by Stuart Russell and others, is to propose a framework that respects the following three principles (criteria) for value alignment:

1. *The machine's purpose is to maximize the realization of human values*. In particular, it has no purpose of its own and no innate desire to protect itself.

2. *The 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 it may never achieve complete certainty.

3. *Machines can learn about human values by observing the choices that we humans make.*

Inverse Reinforcement Learning

Reinforcement Learning focuses on computing optimal behaviors given a reward function

In inverse reinforcement learning, one does the opposite: **observe optimal behaviors** and try to compute the reward function that agent is optimizing.

Hence a good strategy for value alignment might be for the AI to observe human behavior, learn the human reward function with inverse reinforcement learning, and behaves according to that function.

Note that we don't want the AI to optimize reward for itself (get coffee for itself) but rather to optimize reward for the human (get coffee for the human)

But two challenges remain:

Cooperative Inverse Reinforcement Learning

1) If a human is being observed, knowing that the observer is learning, human likely to act differently – highlighting what are common pitfalls/mistakes

(the human should perhaps explain the steps in coffee preparation, show robot where backup coffee supplies are and what to do if the coffee pot is left on the heating plate too long, while the robot might ask what the button with the puffy steam symbol does)

2) The AI has to both learn its goal (maximise the human reward) **and** take steps to accomplish it.

Inverse reinforcement learning does not allow either of the above. Hence **cooperative inverse reinforcement learning** (CIRL) formalizes value alignment as a game with two players.

Cooperative Inverse Reinforcement Learning

The human knows the reward function, the AI does not, and the AI's payoff is exactly the human's actual reward.

That is to say, not only learning what the human's reward is, but acting to optimise his reward for the human

The AI therefore both learns and acts

For the AI to learn the human's reward and act to optimise it, it will strategically act to get clarification (e.g. ask questions). In turn, since the human knows the AI is trying to help, the optimal cooperation strategy for the human will involve teaching behaviours

Invited Talk Today

6pm in the ***SAFRA LECTURE THEATRE, KINGS Building***

Speaker: Jan Leike

Title: Reward Modeling

Abstract:
We want to apply machine learning to help us solve increasingly complex real-world challenges. However, the real world does not have built-in reward functions and designing such reward functions is difficult in part because the user only has an implicit understanding of the task objective. This gives rise to the agent alignment problem: *How can we create agents that behave in accordance with the user's intentions?* In this talk I will explain how we plan to address the agent alignment problem with *reward modeling*: learning a reward function from interaction with the user and optimizing this function with reinforcement learning.

Bio:
Jan is a Senior Research Scientist at DeepMind where he works on the agent alignment problem. He holds a PhD in computer science from the Australian National University in theoretical reinforcement learning. Before joining DeepMind, he was a postdoctoral researcher at the University of Oxford. Jan's research interests are in AI safety, reinforcement learning, and technical AI governance. (https://jan.leike.name/)