Lecture 12:

Ethics and the Future of Intelligence

COMP90087 The Ethics of Artificial Intelligence

Matthew Farrugia-Roberts

The University of Melbourne

May 23rd, 2024

Overview

Lecture 12: Ethics and the future of intelligence

Timeline: Lecture now–13:00, 5 minute break, resume lecture 13:05–13:55.

Four questions:

1. *Can* society build a superintelligent AI system?
2. *What if* society builds a superintelligent AI system?
3. *Should* society build a superintelligent AI system?
4. *Will* society build a superintelligent AI system?

This lecture is not examinable.

Content Warning

In this lecture we will take seriously certain confronting possibilities:

- **Negative outcomes:** Global catastrophe or human extinction from AI.
- **Hopelessness:** Bad things might happen even if we try to stop them.

Thinking about these topics can be psychologically intense—similar to, for example, thinking about **climate change**.

It is worth having this discussion, because there is still time for society to influence the development of AI technology in a positive direction.

However, depending on your individual experience, this discussion may raise difficult feelings such as distress.

Support Resources (University of Melbourne)

Please note that the University has resources available to support you.

1. **Free self-directed LMS module** "The Sustainable Self" can help you learn general skills to face challenging topics such as these.
2. **Free personal counselling appointments:** book an appointment for same day/next day support from the University's Counselling & Psychological Services team.
3. **Urgent free 24/7 mental health support** from the University Mental Health Crisis Support Service. Phone and SMS numbers listed online along with other crisis support services.

Edit to add: for non-students see resources at HelpGuide.org.

Lecture Outline: Four Questions

Can Society Build a Superintelligent AI System?

What If Society Builds a Superintelligent AI System?

Should Society Build a Superintelligent AI System?

Will Society Build a Superintelligent AI System?

Conclusion

Narrow Artificial Intelligence

Most of today's AI systems:

- Are designed to perform a single task to a human (or superhuman) level of performance.
- Operate in highly restricted (digital/virtual) settings with specially prepared inputs.

Examples:

- Image classifiers, facial recognition systems.
- Automated decision-making systems in finance, law, medicine.
- Board game and video game AI systems.

The story of AI so far is a story of designing narrow systems that can cope with more and more complex tasks.

General Artificial Intelligence (AGI)

Imagine a future AI system that is:

- Capable of performing a broad range of tasks (human-level breadth and human-level performance).
- Able to operate in an open-ended environment with flexible inputs (like the real world or a complex virtual world).

Proto-example:

- Large language models (can perform a large range of tasks, has superhuman breadth of intuitions, performance is fragile and unfocussed, operates primarily in linguistic environments or with trained embeddings).

So-called "AGI" is a long-running goal of the field of AI research.

Is AGI Technically Possible?

Potential paths to general intelligence:

- **We're nearly there?** Perhaps current techniques, scaled up, will get us to AGI?
- **Whole brain emulation?** Perhaps a neuron-by-neuron implementation of a human cognitive architecture would be possible?
- **"Reward is enough"?** Maybe a complex open-ended environment and some simple objective, plus enough brute-force trial and error, will lead to something intelligent?

There is much debate about *how close* we are to AGI today, but most AI researchers agree that *at some point* we will be able to develop AGI.

(Not to say that this is an unbiased sample that has never found out things are harder than they seemed.)

Superhuman Artificial Intelligence (Superintelligence, ASI)

Is human-level breadth and performance the limit?

A **superintelligence** is "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest" (Bostrom).

Analogies:

- Greatly exceeding the performance of the most intelligent humans (in every domain).
- Even exceeding the collective performance of humanity as a whole.
- Including non-intellectual domains such as motor control, social intelligence, emotional intelligence, creativity, leadership, rhetoric.
- Don't think "Einstein vs. village idiot", instead think "human vs. insect."

Is Superintelligence Technically Possible?

Assuming AGI is possible, is superintelligence possible?

Some reasons to think that superintelligence might follow AGI:

- **Speed improvements:** We can probably speed up AGIs to enable greater rates of cognition than humans.
- **Quantity improvements:** We can probably replicate AGIs into populations of AGIs.
- **Qualitative improvements:** We might discover fundamentally better-than-human cognitive approaches while searching for AGI.
- **Recursive improvements:** Progress might accelerate after AGI, since AI systems could then contribute to iteratively improving AI systems.

What is the new equilibrium? Humans appear to be at an approximate equilibrium of intelligence, but this is due to evolutionary constraints.

Intelligence versus Consciousness

Sometimes people ask a different question: "Can society build an AI system that is conscious, that is, it has a subjective experience?"

We do not understand consciousness, even in humans. The connection to AGI/ASI is uncertain:

- It may be possible to create a general/superhuman AI system without creating a conscious AI system.
- Or, maybe intelligence is connected to consciousness, and to create AGI/ASI we must create a conscious AI system.

However, creating consciousness *may not require understanding consciousness.*

Thus, from a technical perspective, machine consciousness seems irrelevant.

(From an ethical perspective, machine consciousness may be very relevant.)

Lecture Outline: Four Questions

Can Society Build a Superintelligent AI System?

What If Society Builds a Superintelligent AI System?

Should Society Build a Superintelligent AI System?

Will Society Build a Superintelligent AI System?

Conclusion

Intelligence is Foundational

Human intelligence is a foundational building block of society.

> *Everything civilization has to offer is a product of our intelligence.*
>
> *Our intelligence is responsible for our civilization. With access to greater intelligence we could have a greater—and perhaps far better—civilization.*
>
> *—Stuart Russell, Human Compatible*

It's true that with our intelligence we have made progress shaping the world to suit us. I would add, the current limitations of human intelligence also influence much of the way society has evolved.

Removing a constraint changes the equilibrium.

How Good Could it Be?

To a first approximation, ASI could transform the world in any way we want.

- Accelerate life/medical science for greater control over individual human physical and mental health, solve illness, aging, genetic limitations.
- Accelerate physical science and engineering for greater control over our natural world and our universe.
- Solve social conflict by developing stable cultures and social institutions according to arbitrary chosen principles.

More intelligence could help solve most current barriers to human flourishing.

Many things we could think to want would suddenly be within reach, and we would only have left to decide what to do with our new power.

How Bad Could it Be?

There are also many possible negative outcomes of ASI society could find itself stumbling into:

- Large-scale economic displacement and disempowerment of human populations or arbitrary concentrations of wealth and power.
- Erosion of social stability by undermining crucial social institutions, such as through false and unverifiable media, or widespread automated fraud and other criminal activity.
- Lock-in of authoritarian political regimes enabled by pervasive surveillance technology or other means.
- Destructive warfare with autonomous weapons or newly-engineered biological weapons, from nations or individual bad actors.
- Loss of control of the ASI leading to human extinction.

Wait, Human Extinction?

Yes, some have argued that human extinction is plausible, and many (though not all) AI scientists take this seriously.

> *Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.*
>
> *—CAIS Statement on AI Risk, 2023*

Proposed general mechanism for existential risk from superintelligent AI:

1. **Loss of control:** Once an ASI is deployed, we won't be able to control it, unless it is specifically designed to be controllable.
2. **Misalignment:** Once an ASI is deployed, it will transform the world in a way we won't be able to survive, unless it is specifically designed to maintain human survivability.

The Control Problem

> *If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere once we have started it, . . . then we had better be quite sure that the purpose put into the machine is the purpose which we really desire and not merely a colorful imitation of it.*
>
> *—Norbert Wiener, 1960*

For most tasks, a system has an *instrumental incentive* to:

- . . . stop anyone from disabling it or changing its task.
- . . . circumvent constraints and flaws in its capabilities.
- . . . devote resources towards the pursuit of its task.

Implication: If the design of our first ASI doesn't work as we intended, we may find ourselves stuck, or facing an adversarial contest with an ASI.

The Alignment Problem

So far, nobody has proposed a design for an AI system that will understand and respect its designers intentions/values, or those of society.

Instead, "AI alignment" researchers have found, for example:

- Human values and related concepts are hard to programmatically define.
- Human values could perhaps be learned from humans, but it is still difficult to define a sound and robust data stream.
- Even if defined or learned, it is hard to get a *learned* AI system to internalise these values completely.
- Alignment becomes difficult to verify for increasingly intelligent systems.

Moreover, any difference in values may become amplified to an extreme degree by a superintelligence.

Extinction becomes a possibility even without attributing malicious values, because humanity could not survive *most* changes to the world.

Lecture Outline: Four Questions

Moral Philosophy To The Rescue?

The design and building of a superintelligent AI system is an ethical dilemma:

- There are (potentially) very serious risks—as bad as human extinction.
- *If* the risks can be averted, there are many potential benefits!
- (Even then, we have to decide *which* potential 'benefits' to realise.)

To navigate this dilemma, we turn to the philosophers:

- Ethical frameworks can offer principles to guide us towards the right path forward.
- (If we can agree on the right philosophical approach. . .)

https://existentialcomics.com/comic/202

Who is the Moral Agent?

"Should society build a superintelligent AI system?" is not a precise enough question for us to apply an ethical framework.

Who is the **moral agent,** the being for whom it would be right or wrong to (contribute to) building a superintelligent AI system?

- Individual executives or politicians?
- Individual researchers or engineers?
- Individual investors or democratic voters?
- Collectives such as corporations, countries, or all of humanity?

Each decision-maker faces a slightly different moral question.

Who are the Moral Patients?

We should also decide whom to consider as **moral patient,** the beings eligible for moral consideration by the moral agent.

- All customers (of a corporation) or citizens (of a state)?
- All living humans?
- Should we also include animals, and/or the environment?
- Should we perhaps include past humans and/or future humans?
- Should we perhaps include future intelligent AI systems themselves?

Depending on the set of moral patients, we might face different considerations.

What Does Utilitarianism Say?

What are the consequences? Roughly:

- There is some probability of lots of positive utility if we build ASI.
- There is some probability of lots of negative utility if we build ASI.

Principle of utility then says: quantify those probabilities and utilities, and proceed if the balance is positive!

Of course, it's not actually binary decision:

- Probabilities and utilities depend heavily on the design.
- The ability to make the right eventual choice may be improved by further research.

Conclusion: Pursue research *and* development aiming to maximises expected utility, including identifying and avoiding risks.

What Does Deontology Say?

What are the relevant duties?

- Ross's **duty of beneficence:** Promote the well-being of others.
- Ross's **duty of non-maleficence:** Avoid harming others.
- Kant's **formula of humanity:** Respect people's autonomy.

Conclusions:

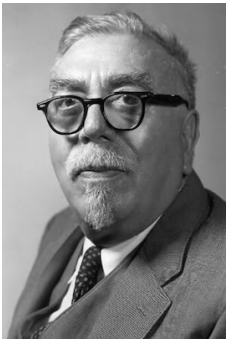
- Perhaps even if it looks like a "good bet" to attempt to build and control ASI, any individual has a duty to *not* unilaterally endanger all of humanity.
- We also have a duty to work hard towards the benefits of ASI, though we should do so *carefully*, anticipating and avoiding risks.

What Does Virtue Ethics Say?

Virtue ethicists invites us to consider a character-based perspective on how to approach the design of ASI.

Some **candidate moral exemplars** in risks from technology:

Norbert Wiener



Yoshua Bengio

Geoffrey Hinton

Helen Toner

Lecture Outline: Four Questions

Incentives for Automating Intelligence

Many agents see strong incentives to pursue marginal improvements in automated intelligence (up to the point of losing control):

- **Corporations** and their executives: market incentives.
- **Nations** and their leaders: economic incentives, military incentives, sovereignty incentives.
- **Investors:** market incentives.
- **Researchers/engineers:** spirit of invention, market incentives.

Exception: Some people may experience negative effects from marginal increases in intelligence as they disrupt and reorganise the economy, however, usually not those in a position to carry out the automation.

The Unilateralist's Curse

Calculated risks: Even if a decision-maker can see risks of loss of control or flaws in their design, they may try to build ASI anyway if:

- they have sufficiently high risk tolerance, and/or
- they are sufficiently optimistic about their approach.

The optimiser's curse: If a large number of groups face this choice, they will naturally have a spread of risk tolerances and evaluations of their approach.

The group with the most optimistic viewpoint and the highest risk tolerance will have a wildly miscalibrated perspective.

The unilateralist's curse: Yet, this group cannot detect that their views are miscalibrated. From their perspective, it makes sense to take the risk!

This increases the likelihood for society to become exposed to this risk.

AI Race Dynamics

Competition incentivises people to cut corners on safety and ethics.

Thought experiment: Suppose you lead an AI research corporation and you have a design for an ASI. Your AI safety team says there is an unacceptably high risk of severe misalignment. Your options:

- Wait until your safety team has a solution?
- Give your safety team 6 months to develop a solution, then deploy the ASI with that solution?
- Fire your safety team and deploy the ASI as-is?

Question 1: What should you do?

Question 2: Does your answer change if you additionally know that your main competitor, who has no ASI safety team and is openly dismissive of risks from ASI, will have an ASI ready to deploy in 6 months?

The Coordination Problem

We face a **tragedy** where incentives and the unilateralist's curse are leading everyone to race toward an outcome no-one involved wants (let alone what is ethically preferred).

We need to **coordinate** between all groups involved in AI development to:

- Remove these harmful incentives.
- Prevent groups from taking unilateral action.

Some avenues for coordination:

- "AI governance"—national and international regulations?
- Doctrine of "self-assured destruction"?

A Further Dilemma

Coordination comes with costs:

- Preventing misuse suggests avoiding "open source technology" whereby safeguards can be more easily removed.
- Non-proliferation of AI technology amounts to concentration of power.
- Arduous regulation makes it hard for new developers to contribute, pushing against inclusivity.
- Defeating the unilateralist's curse may require invasive surveillance and restrictions of freedom.

Bootleggers and Baptists: This makes worrying about existential risk sound like a great business strategy for AI companies...

This makes finding an *ethical* path forward *that much harder!*

Ethics versus Safety versus Accelerationism

There is a furious debate raging in the public sphere between various camps:

- **AI safety-ists**, concerned about risks from *future* AI systems.
- **AI ethics-ists**, concerned about risks from *present-day* AI systems.
- **Humanist accelerationists,** dismissive of risks, therefore compelled by the potential benefits to charge ahead.
- **Posthumanist accelerationists,** pursuing the advancement of intelligence itself, whether or not humans survive.

Perhaps surprisingly, there is an implicit coalition between the accelerationist camps, despite a fundamental clash of values.

Perhaps surprisingly, there is *bitter rivalry* between the first two camps, despite a fundamental concern for risks from AI.

Lecture Outline: Four Questions

Can Society Build a Superintelligent AI System?

What If Society Builds a Superintelligent AI System?

Should Society Build a Superintelligent AI System?

Will Society Build a Superintelligent AI System?

Conclusion

Four Answers?

1. ***Can* society build a superintelligent AI system?**
   - It seems plausible, but there may be unforeseen barriers.
   - Timelines unclear.
2. ***What if* society builds a superintelligent AI system?**
   - It will completely transform society, for better or worse.
   - It's plausible (not certain) that, without careful design, the outcome could be as bad as human extinction.
3. ***Should* society build a superintelligent AI system?**
   - Yes, we must do so, for the benefits!
   - ... *if* we can do it safely, carefully managing serious risks.
   - ... *and if* we can do this without compromising ethics in the mean time.
   - ... *and if* we can design it in an inclusive manner.
4. ***Will* society build a superintelligent AI system?**
   - We would appear to be racing towards this goal about as fast as possible.
   - Slowing down requires solving a difficult coordination problem.