The Philosophy and Ethics of Artificial Intelligence

Dr. Sanjay Modgil

N7.08 Bush House

Office hours: 11-1 Mondays

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

- Filter Bubbles and Echo Chambers
- Surveillance Capitalism
- The Impact of AI on what it is to be Human
- AI and Work
- Conscious AI

The Vision of the Internet

The early pioneers of the internet and web had a vision

Access to the world's knowledge would expand peoples' horizons, making them more knowledgeable, aware of facts, exposing them to different views from their own, and encouraging debate and discussion.

Where did it all go wrong ?

E.g Sir Tim Berners-Lee described renaming of UK Conservative Party Twitter account as a fact checking body as "impersonation".

He has launched global action plan to save web from political manipulation, fake news, privacy violations and other malign forces that threaten to plunge the world into a "digital dystopia".

Filter bubbles and Echo Chambers

People are increasingly receiving their news online through non-traditional sources, such as Facebook, Google, and Twitter that use filtering algorithms that tailor individuals' news feeds. This method of curating content has replaced the function of the traditional news editor.

The filtering algorithms ***personalise*** the content you see – what you've already seen is assumed to be what you desire to see, so lets fulfill your desires and give you more of what you want to see and filter out what you don't want to see.

➔ **filter bubble** is a state of intellectual isolation that can result from personalized searches when a website/social media platform uses **filtering algorithms** to selectively guess what information a user would like to see based on information about the user, such as location, past click-behavior and search history.

Filter bubbles

Feedback mechanism: when it comes to news/opinion you get more and more information that reinforces your views, and you become separated from information and opinions that disagrees with their views

Our views become more rigid (and possibly even more extreme = “belief polarisation”) and hence less likely to be changed by debate, reason and argument.

We become shielded from diverse perspectives crucial for creating well-informed citizens who are tolerant of others’ views and ready to have our views changed.

Echo Chambers

Issue worsened by echo chamber phenomenon

Echo chambers can exist without ”help” from filtering algorithms
They refer to overall phenomenon by which individuals are exposed only to information from like-minded individuals, while filter bubbles are a result of algorithms that choose content based on previous online behavior.

- one source of information makes a claim, which many like-minded people then repeat, overhear, and repeat again (often in an exaggerated/distorted form) until most people assume that some extreme variation of the story is true.

- Repetition in online “tribal” communities ➔ participants find their opinions constantly echoed back to them, which reinforces their opinions

- Community members feel more confident that their opinions will be more readily accepted by others in the echo chamber

=> Mutual reinforcement (and possibly polarisation) of beliefs.

Fragmentation of World Views

Social networking communities are powerful reinforcers of rumors because people trust evidence supplied by their own social group, more than they do the news media

Online social communities therefore become fragmented when like-minded people group together and members hear arguments in one specific direction.

In certain online platforms, such as Twitter, echo chambers are more likely to be found when the topic is political in nature compared to topics that are seen as more neutral

To solve problems we need a shared understanding of how things are

The Infopocalypse (Information Apocalyspse) is Nigh !

- Multiple non-traditional news sources on social media, combined filtering algorithms and echo chamber phenomenon => polarisation, fragmentation and entrenchment of beliefs

- Effects dangerously amplified by multiple sources peddling fake news/alternative facts in post truth world in which there is diminished respect for truth

- Used by groups within states and by states (notably Russia but others as well) to undermine societal cohesion and democratic processes

The Infopocalypse is Nigh !

- But we are now also seeing rapid technological advances in development of **deep fakes** (synthetic media generated by machine learning AI) to generate images and videos (e.g. a totally realistic video of me can be generated in which I am seen denying that the COVID crisis is real)

- This is a new field −there will be legitimate uses (eg advertising) but can far more effectively create fake news, misinformation for disruption of societies and political institutions

Deepfakes: The Coming Infocalypse. Nina Schick

What helps explains these phenomena ?

- What may help understand the phenomena of filter bubbles and echo chambers is insights into how human reasoning evolved

- D. Sperber and H. Mercier's *Argumentative Theory of Reasoning*

The Enigma of Reason: A New Theory of Human Understanding
Dan Sperber and Hugo Mercier. 2017. Penguin Books.

Why do humans reason? Arguments for an argumentative theory. Hugo Mercier and Dan Sperber. Behavioural and Brain Sciences (2011)34, 57–11

The Argumentative Theory of Reasoning

In a nutshell:

- Sperber and Mercier propose that instead of having a purely individual function, reasoning has a social and, more specifically, argumentative function. The function of reasoning would be to find and evaluate arguments in dialogical contexts; that is, to argue with others.

- Contrast with classical *Cartesian* individualistic view of reasoning - reasoning evolved to critically examine our beliefs so as to discard wrongheaded ones and thus create more reliable beliefs, and so make better decisions

- But plenty of psychological evidence demonstrating failure of reasoning in decision making (e.g. *Thinking Fast and Slow*. Daniel Kahneman)

The Argumentative Theory of Reasoning

- Humans became more and more social, collaborating to hunt, trade, raise children, etc. To collaborate requires that we **communicate**.
- However, for communication to be possible, listeners have to have ways to discriminate reliable, trustworthy information from potentially dangerous information— otherwise speakers may abuse listeners through lies and deception.
- That is: to avoid being misled and possibly manipulated into acting against one's self-interest/survival, it is evolutionarily advantageous for a listener to exercise 'epistemic vigilance' (especially when receiving information from speakers in whom the listener does not have a high degree of trust and when the information does not cohere with what she believes).
- This vigilance is exercised by evaluating reasons (i.e., arguments) for the received information, and looking for counterarguments, before accepting the received information

The Argumentative Theory of Reasoning

- In turn, it is to the advantage of the speaker that she produce arguments supporting the information being communicated, in order that it be accepted.
- Reasoning therefore evolved to produce and evaluate arguments in these communicative (dialogical) settings, therefore increasing quantity and quality of the information humans are able to share, by allowing speakers to argue for what they claim, and by allowing listeners to assess these arguments and look for counter-arguments.
- The theory explains why when people reason alone they look for arguments/evidence to confirm their beliefs, and fail to look for counter-arguments/evidence against their beliefs. This is the well known **confirmation bias** extensively studied in psychology.

The Argumentative Theory of Reasoning

The theory also explains **reason based choice**

- According to the reason-based choice framework, people often make decisions because they can find reasons to support them. These reasons will not neccessarily favor the best decisions or decisions that satisfy some criterion of rationality, but rather decisions that can be easily justified and are less at risk of being criticized.

- Again, this is predicted by Sperber and Mercier’s theory. Decision makers are disposed to make decisions using reasons/arguments that they anticipate having to communicate to others and defend from criticism from others. These are the arguments that can be easily justified and are less at risk of criticism, rather than arguments that rationally support the decision.

Back to Filter Bubbles and Echo Chambers

Before the web, internet and social media our inclinations (evolutionary dispositions) to preferentially seek out arguments/evidence in support of our beliefs, relied on conventional media and interactions with others

But now, social media puts us in touch with vastly more people with similar beliefs, and the filtering algorithms vastly amplify the extent to which we are both fed with confirmatory evidence/arguments for our beliefs, while at the same time shielding us from challenges and opposing views and evidence !

Reason based choice and the confirmation bias is massively enhanced - our beliefs and decisions become more entrenched/rigid.

We have an example of AI amplifying cognitive biases of humans and exploiting humans' desire to belong to tribes (identity groupings) both of which arguably hold back moral progress and well being

But not all is lost …we perform better when engaging in dialogue

- If reasoning evolved so we can argue with others, then reasoning should give better results in groups than alone.

- In dialogical settings a participant (speaker) is not only disposed to put forward arguments that support their beliefs/decisions given the presence of listeners (as predicted by the evolutionary theory), but is also now no longer blinded to counter-arguments, and challenges which are provided by other participants (listeners), as predicted by the evolutionary theory.

- And indeed evidence shows that it does ! When the performance of groups and lone individuals in reasoning tasks is compared, groups do much better— sometimes dramatically so.

- With more arguments at hand, better decisions may be made, critical thinking skills are improved, and exposure to different views/ opinions **may** prevent isolation of people in belief bubbles.

Dialogical Scaffolding for Human-Human Reasoning

- In lecture on reasoning and communication we saw how argumentation based models of dialogue can support joint (distributed reasoning amongst multiple agents)
- These models can also support human-human dialogue by guiding humans as to how to rationally engage in argument and counter-argument.

E.g.

- Ag1: “Tony Blair is no longer a public figure, and the information about his affair is not in the public interest, so the information should not be published”
- Ag2: “But Blair spoke on the radio about the importance of marriage vows”

AI – “which premise do you disagree with ?”

- There are examples of applications that support ‘deliberative democracy’ – the idea that we should deliberate more and be encouraged to engage rationally in political debate is one of the most important ideas in political science –
- AI Applications can support rational exchange of arguments and even contribute and provide access to supporting and relevant evidence

Dialogical Scaffolding for Human-AI Reasoning

- These models can also support human-AI dialogue applications for use in educational contexts

E.g.

- Studies show that over 50% of junior doctors' acquisition of clinical reasoning skills to decide amongst alternative (i.e.,conflicting) diagnoses and treatment options, occurs on **ward rounds**.
- Teachers engage students in dialogue, challenging their assumptions, suggesting alternative diagnoses and treatment options, and suggesting hypothetical scenarios that may, for example, prompt the student to propose additional or alternative treatments
- However demands on teachers' time presents significant barrier to such learning
- E-Clinic application in which AI plays role of teacher, engaging students in dialogue and improves medical reasoning skills.

- Other educational contexts, eg. Politics class, Philosophy class, This class …

AIs that Argue and Debate

- AI scraping the web for **arguments**
- **Computers that can argue will be satnav for the moral maze. New Scientist,** Issue 3090 published 10 September 2016

Read more: https://www.newscientist.com/article/mg23130900-700-rage-against-the-machine-why-computers-need-to-argue-with-us/#ixzz64PhTHUmB

- IBM **Project Debater**

The latest IBM grand challenge (after Deep Blue and then IBM Watson)

Project Debater is the first AI system that can debate humans on complex topics. Project Debater digests massive texts, constructs a well-structured speech on a given topic, delivers it with clarity and purpose, and rebuts its opponent. Eventually, Project Debater will help people reason by providing compelling, evidence-based arguments and limiting the influence of emotion, bias, or ambiguity.

Read more: https://www.research.ibm.com/artificial-intelligence/project-debater/

-

Back to Filter Bubbles and Echo Chambers

- One might think that exposing people (in social media contexts) to more opinions and arguments that challenge their beliefs/opinions will help solve the problem

- But evidence that more exposure also leads to more rigid adherence to beliefs – especially idealogical beliefs about identity/politics

- Crucial difference with educational settings is that in social media context peoples' intentions are not to get to the truth of the matter/have their views challenged. Instead their intention is to seek news/opinion that support their views and amplify sense that they belong to a tribe.

Hence they tend to dismiss opposing views/evidence, and the more they do so, the more their views are confirmed.

Back to Filter Bubbles and Echo Chambers

- Tackling the problem of filter bubbles and echo chambers requires more than just increasing exposure, but also **changing the way we interact with information**.

- **From search to dialogue**: early educational interventions – children from a young age using **dialogue engines rather than just search engines** – exposure to arguments and dialogue = developing critical reasoning skills.

There may be hope given recent critiques (Celia Heyes and Catarina Noaves) of Sperber and Mercier, that they underplay extent to which social interactions (rather than evolution) account for biases (redrawing the line in the age old nature/nurture debate)

Dialogical Scaffolding for Moral Reasoning

- Arguably any moral/ethical decision can be understood as a utility maximizing problem, where (in the case of utilitarianism) utility equates with **happiness/well being**.
- Happiness/well-being based on subjective/conscious human experience
- Hence complex/moral ethical decisions will need to incorporate human input
- In fact, moral decision making – especially for novel ethical issues that have no precedent and for which humans are still undecided - **can be improved** by incorporating future AIs' superior **epistemic** reasoning (what is the case) and **causal reasoning** about the consequences of actions, and integrating this reasoning with human inputs relating to human values/ preferences that are rooted in subjective well being
- ➔ Integrate human and AI reasoning through argumentation based models of dialogue specialized for reasoning about moral/ethical issues (see

Example Debate

- See *Many Kinds of Minds are Better than One:Value Alignment Through Dialogue. S. Modgil*

  (https://nms.kcl.ac.uk/sanjay.modgil/ExtendedAbstract.pdf)

  for a short example of how an AI might jointly reason with humans to help decide a policy that a autonomous vehicle should use when choosing to endanger the life of the driver or pedestrians (as in the moral machine experiment https://www.moralmachine.net/)

Surveillance Capitalism

Surveillance Capitalism

- *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. Professor Shoshana Zuboff

- History of Capitalism: What was outside the market became commodities - things to bought and sold for a profit

  - nature commodified as real-estate
  - human activity commodified as labour
  - education
  - **Now private human experience - behavioural data - bought and sold on the market**

The Logic and Birth of Surveillance Capitalism

- Basic and powerful idea: the more data about your beliefs, desires, goals, behaviours, etc etc the more one can predict what you will (want to) do, and so manipulate you into doing things that serve interests of companies/corporations who want to profit from you, and not necessarily the interests of your well being/happiness (sounds familiar ?)

- Google started of not wanting to advertise but then dot com crash in early 2000s – they realised they had all this surplus data that they could use to make money. Also 9/11 in 2001 meant less governmental restrictions on monitoring data !

- So initially surveillance capitalism born and raised to serve online targeted advertising : Google data about your searches/clicks used to predict what you like/future behaviours and then sold to companies to enable targeted advertising. Indeed, click through rates dramatically increased !

The attention economy

- Advertisers need your attention

- The **exposure** to targeted advertising is being increased by the planned and strategic development of persuasion technologies designed to keep your eyes and attention fixed on sites and social media where adverts are being presented

- E.g. the extent to which social media manipulates fears and anxieties of teenagers – their desire for validation, recognition, to be **liked** by their peers

Mining Behavioural Predictions – beyond click/search data

- Now private human experience - not just behavioural data that is about what we desire and so are more likely to buy online – is more and more being gathered from **every** aspect of our lives and input to machine learning algorithms that produce predictions of our purchasing **and other** (e.g. voting) behaviours.

- The alarm beside your bed rings, triggered by an event in your calendar. The smart thermostat in your bedroom, sensing your motion, turns on the hot water and reports your movements to a central database. News updates ping your phone, with your daily decision whether to click on them or not carefully monitored, and parameters adjusted accordingly. How far and where your morning run takes you, the conditions of your commute, the contents of your text messages, the words you speak in your own home and your actions beneath all-seeing cameras, the contents of your shopping basket, your impulse purchases, your speculative searches and choices of dates and mates – all recorded, rendered as data, processed, analysed, bought, bundled and resold like sub-prime mortgages.

- New marketplace that trades in these predictions/behavioural futures.

From the online world to the real (physical) world

- Health, insurance, education, retail, sectors etc all purchasing behavioural predictions to achieve profit

- Now political sector using profiles and predictions to target voters and manipulate choices - e.g. Cambridge Analytica – although political scientists sceptical about effectiveness of micro-targeting, but in the future may become more effective !

- *Pokemon Go* – developed by a subsidiary company of Google took online manipulation to another level ➔ manipulation of behaviours/choices in physical rather than online world by placing characters to collect in shops that paid for this, so increasing number of customers visiting their shops

What’s the Harm ?

Value of privacy undermined – but values are question begging – what are the consequences ? After all, you might think, great, I’ll happily trade my privacy for a more personalised experience !

- But as we see more and more behavioural data collected – opportunities for control and manipulation of our behaviours/choices increases, for example our political choices ➔ undermining democracy.

- Less opportunity for “serendipity” = chance discovery of new things/opinions that we might like/challenge our current opinions.

- Above relates to fundamental sense that we have free will. We have reviewed arguments as to why free will is an illusion; that we think of ourselves as having free will because we that’s what it *feels like* when we make choices.

  But arguably, the sense that we are free to choose also depends on the degree to which we are conscious of the fact that our available choices are not manipulated by others – others who might not have our best interests at heart. The sense that we are not being manipulated, that we are autonomous is an essential component of our well being

What's the Harm ?

- Leaked Facebook document in Australia addressed to business customers, describing how it uses data on 6.4 million young adults and teenagers – simulate and predict individual and group emotional states – so that one can identify when they need confidence boosts and therefore when most vulnerable to specific combination of ads, cues and nudges. It can tell, it boasts, when a young person feels anxious, stressed, stupid, inadequate, like a failure, and so send an add saying here is a leather jacket at discounted price – **buy and it will make you feel great**.

https://www.independent.co.uk/news/media/facebook-leaked-documents-research-targeted-insecure-youth-teenagers-vulnerable-moods-advertising-google-a7711551.html

- We are manipulated into thinking that human well being/happiness/flourishing, depends on material acquisitions (at least to a much greater extent than is supported by scientific research) because this serves profit motive of corporations/capitalism

What’s the Harm ?

- Surveillance capitalism is the ultimate way to manipulate us into participating in/being addicted to retail therapy. It predicts when you are feeling crap and then targets you – buy this ! – you buy and get a dopamine hit!

So in summary, we again see examples of AI being used to amplify aspects of human behaviour that evolved for good reason in different contexts – the desire to be liked and accepted by our peers, and the desire and acquisition of material things – worked well for our distant ancestors who lived in smaller communities when reputation really did matter and resources were scarce

What’s the Harm ?

So in summary, we again see examples of AI being used to amplify aspects of human behaviour that evolved for good reason in different contexts – the desire to be liked and accepted by our peers, and the desire and acquisition of material things – worked well for our distant ancestors who lived in smaller communities when reputation really did matter and resources were scarce

But in modern societies when resources are plentiful and the community more and more global, these instincts/innate desires are less relevant/ useful when it comes to enhancing our well being and to leading a flourishing life

And yet the AI fuelled feeding of these desires turns us into addicts with plentiful all-consuming opportunities for that next short lived dopamine hit that comes from a like, a purchase, and all in the furtherance of the goal of maximising profits. while undermining what it is to experience sustainable well-being

The Impact of AI on what it is to be Human

The Impact of AI on what it is to be Human

- So far in this lecture we have looked at how AI might amplify existing human dispositions and tendencies, to belong to tribes and groupings of shared identity, to have our beliefs and opinions confirmed and amplified back to us, to be liked and get recognition, to acquire, purchase, buy …. and how amplification of these tendencies may not be conducive to our well-being/flourishing

- But AI may also **qualitatively** change what it is to be human, both for the better and for the worse

Artificial Intelligence and the Future of Work

- As of July 2019 mixed reports as to the effects of AI on jobs

Net loss in jobs

- In 2013 a pair of Oxford academics, C. Frey and M. Osborne, estimated that 47% of American jobs are at high risk of AI automation by mid-2030s.
- McKinsey Global Institute: between 40 &160 million women worldwide may need to transition between occupations by 2030, often into higher-skilled roles. Clerical work, done by secretaries, schedulers and bookkeepers, is an area especially susceptible to automation, and 72% of those jobs in advanced economies are held by women.
- Oxford Economics: up to 20 million manufacturing jobs worldwide will be lost to robots by 2030.

as well as many others ….

Artificial Intelligence and the Future of Work

Net gain in jobs

- World Economic Forum: automation will displace 75 million jobs but generate 133 million new ones worldwide by 2022

- Gartner: AI-related job creation will reach two million net-new jobs in 2025.

- McKinsey Global Institute: worldwide, with sufficient economic growth, innovation, and investment, there can be enough new job creation to offset the impact of automation, although in some advanced economies additional investments will be needed to reduce the risk of job shortages. In the US, there will be net positive job growth through 2030.

as well as many others ….

Artificial Intelligence and the Future of Human Employment

But notice that these predictions focus on the relatively short term future

Once we approach AGI and beyond, it is likely that there will be a net loss in traditional jobs

What are the ethical implications ?

- 3 percent of all working American are drivers of some sort and so are vulnerable to being made unemployed by AI. On the other hand, AVs will lead to less injuries and deaths

- In the long term, some people argue that AI will result in great economic benefits – we will then be able to *afford universal basic income* – a minimum payment that enables humans to afford basics. On the other hand, numerous studies demonstrate the importance of work to providing people with meaning/purpose and hence well being. But then can people could work on more creative activities !

Artificial Intelligence and the Future of Human Employment

What are the ethical implications ?

- Cognitive decline: As humans become more dependent on AI to do tasks, they exercise relevant cognitive skills less, which may therefore ultimately decline due to lack of use (e.g. with GPS and sat nav, who now can read a map ?)

- Evidence also that repeated engagement with certain kinds of cognitive tasks have benefits that apply to other cognitive tasks

Conscious AI and its Impact on Human Relations

In the long term, suppose AIs become conscious and so experience suffering, pain, pleasure … ?

Then they should be regarded as agents with moral status and awarded *rights* and treated the way we do (should !) treat other conscious beings (note that time dilation will mean that an AI who suffers over a minute of human subjective time will suffer over far far longer periods of AI subjective time !)

How could we know with absolute certainty that an AI experiences feelings (is conscious) ? As well as information theoretic and architectural indicators (e.g. *IIT*), we will rely heavily on our functionalist intuitions – **if it behaves as if it is conscious then It is conscious**

In fact, many studies show that humans (especially when they are young) require very few behavioural cues in order to anthropomorphise (treat as if human) and so assume that intention, desires, feelings, consciousness is present.

Conscious AI and its Impact on Human Relations

Children playing with dolls will quickly anthropomorphise – e.g. if doll's eyes following child around room

Also, there **may** be a strong moral motivation to make AI robots more human like, so that they are more easily integrated into human culture and so interact with humans and learn/align with our values (as do young children when growing up)

But it is highly likely that there **will** be strong commercial motivation to make AI robots more human like in their speech, gestures, etc and so we will inevitably treat them as if they were conscious.

For example, robots for elderly care will be more readily accepted by elderly if they are human like, and show care, concern, empathy …

Sex robots will be more desirable if they are made to be more human like.

Conscious AI and its Impact on Human Relations

But then how will our relations with these robots we consider to be conscious, affect how we relate to other humans ?

1) We will treat conscious robots as slaves (elderly care, servants in homes etc). This might erode our moral objection to treating humans as slaves (as means to ends) !

2) We will treat conscious sex robots as mere objects of sexual gratification. This might mean we are more ready to objectify other humans as mere instruments for our sexual gratification (just as internet porn is having negative effects on humans' ability to form sexually satisfying respectful and meaningful relationships).

3) If we outsource care to robots, to what extent will we become less caring, to what extent will we preserve values of caring for elderly parents, disabled relatives etc (compare with possible decline in cognitive capacities when AIs start taking on cognitive tasks)

Conscious AI and its Impact on Human Relations

4) Evidence that young children test out, to a limited extent, their social relationships with dolls. But as robotic dolls become more lifelike, more of this testing out of how children socially relate, will happen with robotic dolls. But choosing a robot as a companion, you don't navigate relationship, you dictate them ! You are the one with all the power to decide the nature of the relationship.

   In these situations children may lose ability to see world through eyes of another, and hence empathy and responsibility (e.g. taking care of a robot pet rather than real pet, kids learn way of being connected while being given permission to only think of themselves)

Artificial Intelligence and Human Ethics/Morality

We've suggested ways in which AI might negatively impact on human relations. How might AI improve human relations ?

We discussed the importance and difficulty of ensuring that the values of future AIs are aligned with human values and that the kind of AI tools enabling AIs and humans to reason together may assist humans in making better moral decisions

- Cooperative Reinforcement Learning in cases where humans behave optimally and it is enough that humans teach AIs, AIs questions humans, and then simultaneously learn reward function (i.e., values) of humans and act to maximise reward for humans.

- For more complex moral/ethical problems with no precedent (past examples to learn from) and when humans themselves don't know the "right" answer (e.g, virtual bliss of VR / AV moral dilemmas), AI could help humans to decide what is right/wrong.

Artificial Intelligence and Human Ethics/Morality

- Another possibility is the use of AI to amplify our capacities for empathy and so help make humans more ethically responsible to others

- Arguably all ethical/moral theories embrace the fact that humans care for others and that widening the circle of concern (from our immediate kin, to the extended family, the village and so on) partly contributes to moral progress (e.g. **impartial** maximization of happiness in utilitarianism).

- Concern and care for others is partly rooted in our capacity to **empathise**: to see/understand/feel the way others do. This is why novels has been characterized as “empathy machines”.

AI virtual reality could serve as the ultimate empathy machines !