Artificial Intelligence Ethics and Well Being

Dr. Sanjay Modgil
Reader in Artificial Intelligence
Reasoning and Planning Group, King's College London

sanjay.modgil@kcl.ac.uk

Ethics and Well-Being (The 'Good')

- Some (hopefully) uncontroversial assumptions:

1. Governance aims at 'impartially' maximising *well-being* (broadly construed)

2. Many human biases/dispositions evolved to satisfy basic conceptions of well-being (food, shelter …) in radically different and harsher environments of distant ancestors living/hunting in smaller groups or tribes whose survival depended on intra-tribe harmony and confrontation with other tribes

3. Larger social groupings, collaboration, cooperation, division of labour, economic surplus etc enabled pursuit of more sustainable and richly satisfying notions of well-being accompanied by the normative goal of impartial maximising well-being

4. But pursuit of these more enlightened culturally evolved notions of well being often in tension with older more 'hard-coded' biases/dispositions

The Evolving Role of Government and Policy

Assumptions (continued)

5. Hence we witness contractarian encodings of the better angels of our nature whose wings would be clipped by our more base instincts and biases…
… if it were not for policy and regulation that ideally facilitates impartial realisation of more enlightened notions of well-being

- For example taxation to help those less well of, those with disabilities, those who have been dealt a bad hand through no fault of their own, taxation to subsidize arts … sugar tax ? …. prohibitions on drug use ? …

AI and Well Being

What does this all have to do with AI ?

AI has the potential to both massively augment and amplify as well as reproduce those ingrained biases/dispositions that may have been adaptive for very different environments, but which now hinder and may even reverse our modern-day progress towards ever more enlightened notions of well-being,

And understanding why and how would be useful in formulating policy and regulation to combat these potential impacts of AI

Artificial Intelligence: Where we are now and where we are headed

- Old fashioned 'Intelligent Design' approach researchers equipped AIs with logics for rational reasoning – a kind of first principles approach ➔ an 'AI winter'

- But Spring is coming ! New machine learning approaches more closely resemble human acquisition of reasoning skills. Learning from experience (data), from rewards for choosing certain actions over others …

- Remarkable recent successes in specific tasks, eg. AlphaGo, AlphaFold, Autonomous Vehicles …. potential to radically transform human society (as we are beginning to see) in which more and more cognitive tasks delegated to often far superior AIs

- Even more so if achieve holy grail of Artificial General Intelligence = **the ability to accomplish *any* goal at least as well as humans (and most likely much better !).**

Artificial Intelligence: Where we are now and where we are headed

- When will we achieve AGI - survey results from experts in AI

10 %	50%	90%
2022	2040	2075

- Once we achieve AGI likely that AGI will recursively improve itself (using ever more massive power, speed and data) leading to an intelligence explosion ➔

- Superintelligence = *any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest* (within 20-30 years of AGI). Even more transformative and potentially an existential threat with crucial implications for policy makers

(*Superintelligence: Paths, Dangers, Strategies*. Nick Bostrom.)

Examples of AI and Ethical Concerns

- Reinforcing the Myth that Consumption Equates with Well Being
- Entrenchment, Polarisation and Fragmentation of Beliefs
- Delegating Decisions – Unreflective Encoding of Human Biases
- ‘Conscious’ AI

Reinforcing the Myth that Consumption Equates with Well Being

- Machine learning algorithms learn what you like and can predict users' moods based on collected data about search, clicks, sites visited …
- Leaked Facebook document boasted that it could predict depression/ anxiety in adolescents – a good time for 'pushers' to target adverts at adolescents when they are most in need of the dopamine release that accompanies a purchase
- Fleeting moments of well being associated with consumption – made sense when resources scare in ancestral environments
- But plenty of evidence. e.g., from positive psychology, that sustainable well-being/happiness not served by striving for and continuing to consume
- AI being used to perpetuate and entrench myth that consumption equates with well being

- Regulate ?

Entrenchment, Polarisation and Fragmenation of Beliefs

- Confirmation bias – focus on confirmatory evidence/opinions supporting our beliefs while ignoring dis-confirmatory evidence/opinions
- *Bias is to be expected* given adaptive account of how human reasoning evolved: to convince others to do your bidding/adopt your point of view – makes sense that speakers seek evidence/opinion that can be used to reason in your favor, while listeners seek to disconfirm so as not to be manipulated – a kind of division of cognitive labor (see: *The Enigma of Reason*. Sperber & Mercier)
- Social Media filtering algorithms filtering out opposing news and views and feeding you with confirmatory news and views are technological incarnations of these evolutionarily acquired biases ➔ massively increasing exposure to, and so entrenching, confirmatory opinions and beliefs
- But it's far worse than that …
- Studies show that exposure to opposing opinions does not suffice to counteract entrenchment.

Entrenchment, Polarisation and Fragmenation of Beliefs

- Adhering to beliefs are like gang tattoos = acts of self-identifying with your tribe (e.g., anti-vacc)

- Made sense when beliefs = shibboleths distinguishing tribal membership (I'm an anti-vaccer – we're part of the same tribe – lets cooperate to protect our tribe's resources from other tribes)

- Increased attention ➔ increased exposure to ads ➔ increased revenue and we attend more to what affirms our tribal identity. Hence …

- ***….social media algorithms are designed*** to expose us to more extreme and polarizing news/views that: a) support our tribal beliefs and b) caricature and distort the beliefs of rival tribes invoking anger and rancor – emotional reactions that engender more attention

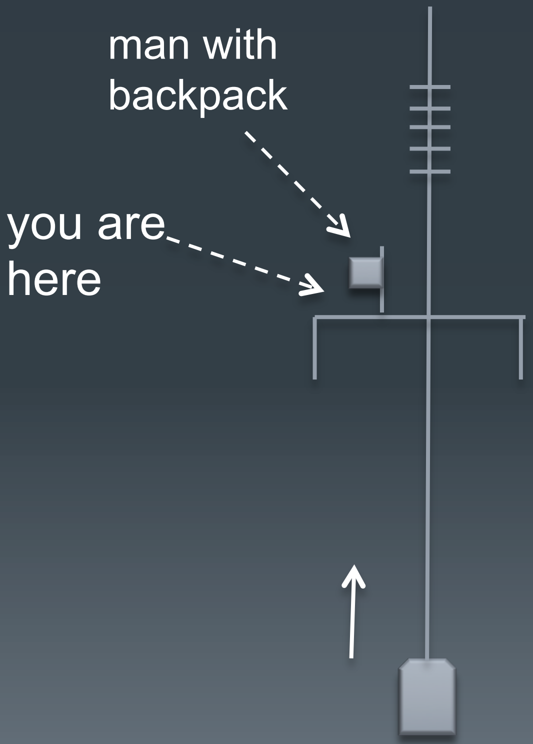
Entrenchment, Polarisation and Fragmenation of Beliefs

- And so people disappear down deeper and deeper rabbit holes of entrenched and extreme beliefs, polarizing and fragmenting world views
- Without a shared understanding/world view solutions to existential problems become harder to agree on and sign up to (the pandemic, climate change, …)
- These effects further amplified by (curated) echo chambers, fake news and deep fakes that can now be developed by AI tools available to anyone ! (see: *Deep Fakes and the Infocalypse: What You Urgently Need To Know*. Nina Schick)
- We surely should be regulating to stem and reverse these existential threats to liberal democratic values

Delegating Decisions

- We will start to delegate more and more decision making to AI and so need to reflect on cognitive biases that do not serve us well, lest we encode these in AI

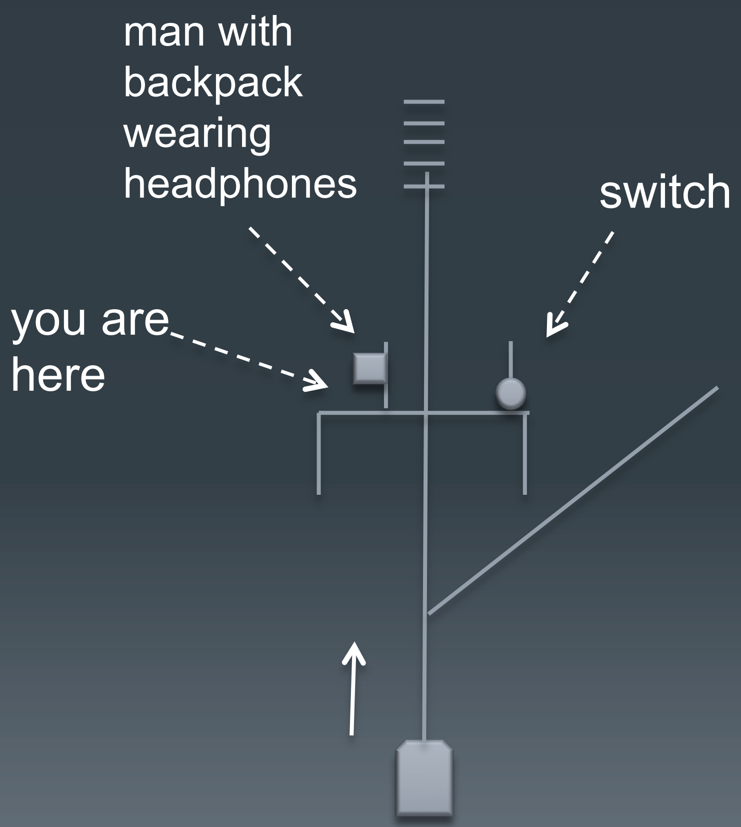
Means to End versus Side Effects



31% would push

Physical Force And Means to End

Means to End versus Side Effects



81% would run to pull switch

Physical Force And Side Effect

Means to End versus Side Effects

- No reason why this distinction should have any moral salience – distinction is morally irrelevant

- Distinction arises because our moral alarms are triggered by immediate cause and effects (means to ends) that our cognitive planning module can keep in view, as opposed to the potentially infinite side effects of our planned actions

- But in international law it is illegal to intentionally bomb civilians to lower the enemies morale (they are used as means) *but legal* to bomb munitions factory knowing full well that civilians will die (side effect / 'collateral damage')

- But why should there be any moral distinction between the two ? And if regulated development of autonomous AI weapons comes to pass, should we not regulate to ensure AI decision making not similarly 'corrupted' ?

Delegating Decision: Autonomous Vehicles (AVs)

- Suppose only way to save life of passenger is to swerve of road and kill ‘gaggle’ of pedestrians

- Car manufacturer justifies algorithm preferentially saving passenger, as a ‘selfish’ car more likely to sell and so increase sales will lead to overall saving of lives as AVs on balance much safer than human driven cars

- But when all cars are autonomous and public cannot choose ?

- If car companies can unilaterally decide what their algorithms do this may lead to a ‘moral sales race’ in which algorithms are more and more ‘selfish’ and so are more likely to be bought by humans with their self-preservation bias, and so resulting in an overall net loss of life

- Regulation required to counteract this commercially exploited self-preservation bias, by enforcing uniform algorithmic decisions that impartially maximise well being

‘Conscious’ Robots

- Well known disposition of humans to readily anthropomorphize
- The more human like in appearance and behaviour the more readily we assume inner conscious experience, including pain, pleasure …
- Sex robots and elderly care robots - commercial imperative to endow with human like appearance and behaviour – in particular behaviour that suggests inner conscious life (she really digs me ! he really cares !)
- We will inevitably come to think of such robots as conscious (of course they’re conscious – how could we have ever thought otherwise !)
- Relations with sex robots may well impact relations with human partners (cf concerns about internet porn)
- We will be treating entities we believe to be as conscious as us, as ‘slave labour’. Will this spill over to corrupt our anti-slavery sensibility ? Will our capacities and instincts for care decline if outsourced ?
- Should we regulate to avoid ‘anthropomorphic trap’ ?

Conclusions: A ‘Joined Up’ Approach

- Examples highlight dangers of AI amplifying and reproducing human biases that are in tension with our modern day conceptions of well being and harmoniously functioning societies

- Examples highlight that a big picture (inter/multi-disciplinary) perspective on the ethical implications of current and future Artificial Intelligence (AI) is needed to inform policy and regulation

- The radically transformative potential of AI means that we need to think about governance now ! (*the precautionary principle*)

- ‘SAGE-AI’ of experts drawing from different disciplines that can be consulted by policy makers ?

Questions ?

Thank you for your attention