Artificial Intelligence and Ethics

Dr BK Murthy

Lecture 3

Introduction

- Ever since the birth of computation with Alan Turing, humans have put high hopes on the power of computers and artificial intelligence (AI).
- The term Artificial Intelligence is Coined by John McCarthy during the Summer workshop at Dartmouth college in 1955
- AI is expected to bring **significant** and diverse benefits to society – from greater efficiency and productivity to tackling a number of difficult global problems, such as climate change, poverty, disease, and conflict.

- AI technologies shape our societies.
- They have an enormous impact on our daily lives.
- At the same time, multiple legal and societal issues have revealed the potential of these technologies to produce undesirable impacts.
- Algorithms can enhance already existing biases.
- They can discriminate.
- They can threaten our security, manipulate us and have lethal consequences.
- People need to explore the ethical, social and legal aspects of AI systems.
- There is a common call for the ethics of AI – meaning how are we to develop and use this technology in an ethically acceptable and sustainable way?

- What are the ethical and moral principles we should adopt and follow?
- In this course, we'll take a look at the ethical issues related to contemporary AI, open up their background in philosophy and give them an interpretation in terms of computer and other sciences.
- The goal of course is to develop skills for ethical thinking.
- The course provides a guide – or a roadmap – on the ethically sustainable design, implementation and use of AI.
- It will introduce you to basic ethical concepts, their theoretical background, and their role in discussion on contemporary AI.

AI/ML

- **Artificial intelligence** is an overall term describing a set of different kinds of techniques to make computers behave in some kind of intelligent fashion.
- There is no agreed definition of AI, but in general the ability to perform tasks without supervision and to learn so as to improve performance are key parts of AI.
- **Machine learning** is a big topic in AI.
- Machine learning is a set of algorithms which by themselves learn to make decisions or to structure data.
- Supervised and unsupervised learning are based on data, while reinforcement learning is where the algorithm uses trial and error to learn to make sequences of decisions.

What is Ethics

- Ethics seeks to answer questions like
  - “what is good or bad”,
  - “what is right or what is wrong”, or
  - “what is justice, well-being or equality”.
- As a discipline, ethics involves systematizing, defending, and recommending concepts of right and wrong conduct by using conceptual analysis, thought experiments, and argumentation.

Subfields of ethics

- **Meta-ethics** studies the meaning of ethical concepts, the existence of ethical entities (ontology) and the possibility of ethical knowledge (epistemology).
- **Normative ethics** concerns the practical means of determining a moral (or ethically correct) course of action.
- **Applied ethics** concerns what a moral agent (defined as someone who can judge what is right and wrong and be held accountable) is obligated or permitted to do in a specific situation or a particular domain of action.

AI Ethics

- AI ethics is a subfield of applied ethics.
- Nowadays, AI ethics is considered part of the ethics of technology specific to robots and other artificially intelligent entities.
- It concerns the questions of how developers, manufacturers, authorities and operators should behave in order to minimize the ethical risks that can arise from AI in society, either from design, inappropriate application, or intentional misuse of the technology.

Main Concerns of Ethics of Technology

- Immediate, here-and-now questions about, for instance, security, privacy or transparency in AI systems
- Medium-term concerns about, for instance, the impact of AI on the military use, medical care, or justice and educational systems
- Longer-term concerns about the fundamental ethical goals of developing and implementing AI in society

From Machine ethics to the ethics of AI

- AI ethics was taken to mean mostly machine and roboethics.
- These cover the study of the ethical codes of artificial moral agents.
- As research fields, they are based on a scenario where machines can one day be responsible for ethically relevant choices, and can even be possibly considered as ethical agents or autonomous moral agents.
- As a comparison, animals are generally not considered moral agents.
- We don’t judge a squirrel’s behaviour as right or wrong, and we don’t assume they have the capacity to know the difference.

From Machine ethics to the ethics of AI…..

- Machine and roboethics span from the development of ethically responsive autonomous vehicles to the design of ethical codes for moral autonomous agents.
- Isaac Asimov,1942 famously proposed “three laws of robotics” that would guide the moral action of machines:
  - A robot may not injure a human being or, through inaction, allow a human being to come to harm.
  - A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
  - A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

- These days, AI ethics is a more general field, and closer to engineering ethics:
- we don’t have to assume the machine is an ethical agent to analyze its ethics.
- Research in the field of AI ethics ranges from reflections on how ethical or moral principles can be implemented in autonomous machines to the empirical analysis on how trolley problems are solved, the systematic analysis of ethical principles such as fairness, and the critical evaluation of ethical frameworks.

Values and Norms

- Values and norms are the basic elements of ethics.
- The concept “value” means, roughly, the degree of importance of a thing or an action.
- Values provide ideals and standards with which to evaluate things, choices, actions, and events.
- In ethics, the focus is primarily on moral values, although other types of values – economic, aesthetic, epistemic (or knowledge-related) – are sometimes relevant morally
- For example, economic factors may play morally significant role, if economic decisions have morally significant consequences to people.

Intrinsic and extrinsic values

- Values can be divided into **extrinsic** (also called “instrumental”) and **intrinsic** values.
- For example, money has extrinsic or instrumental value. Money is valuable only because one can use it for other things, such as to provide better medical care for the people.
- These things, in turn, may be good for what they lead to: for example, for better health.
- And those things, in turn, may be good only for what they lead to – for example the better quality of life.
- Intrinsically valuable things are typically “big moral values” – happiness, freedom, wellbeing.
- These are things that are good as they are.
- For some, they also explain the “goodness to be found in all the other things”.

Norms

- Norms are value-based principles, commands and imperatives – such as the sets of AI guidelines.
- They tell what one should do, or what is expected of someone.
- Norms may be prescriptive (encouraging positive behavior; for example, “be fair”) or
- Proscriptive behavior (discouraging negative behavior; for example, “do not discriminate”).

Type of Norms

- Some norms are merely statistical regularities:
- one notices that many computer scientists tend to wear black T-shirts.
- Regulations imposed by the authorities
- Some norms are social norms;
- They tell what people in a group believe to be appropriate action in that group.
- Regulations imposed by the authorities so that your actions should not cause inconvenience to others

- Moral norms are prescriptive or proscriptive rules with obligatory force beyond that of social or statistical expectations.
- For example, “Do not use AI for behaviour manipulation” is a moral norm.
- Norms may also be legal norms.
- Importantly, a legal norm may not be a moral norm, and vice versa.
- Simply, the fact that “X is a law” does not make it a moral principle.
- Instead, one can always ask: “Is this law morally acceptable or not?”

Hume´s guillotine: Facts, value and norms

- Normative claims do not describe how the world is.
- Instead, **they prescribe how the world should be**.
- This is, they imply “ought-to” evaluations, in distinction to sentences that provide "is" types of assertions.
- For instance, a sentence "This machine-learning system is a black-box system" is descriptive,
- while a sentence “Machine-learning systems should be transparent" is normative.

- Importantly, facts do not dictate our norms.
- As Scottish philosopher, David Hume (1711–76) states it, one should not make normative claims about what should be, based only on descriptive statements about what is.
- This does not mean that facts do not take any part in our moral consideration, but that you cannot get from an “is” to an “ought” without the use of some purely normative value statement along the way.
- This principle is known as “Hume´s guillotine”.

Hume´s guillotine Principle

- Hume´s guillotine principle states that moral norms or claims cannot be justified only by appealing to facts.
- As Hume remarks, one cannot derive the “ought from is”.
- For example, the fact that “there is a biased data set” does not alone imply that the data should (or shouldn’t) be biased.
- Instead, moral attitudes depend on other ethical considerations and preferences, not just mere facts.

- Why are we concerned with the issue of biased data?
- Well, the problem clearly is not the fact that there are biased data.
- The real problem is that biases may enhance discrimination.
- Importantly, Hume´s guillotine does not claim that facts don’t matter. They do.
- The point is that facts (alone) don´t solve ethical problems.
- Instead, ethical problems require genuinely ethical discussion, too.

A Framework for AI Ethics

- Traditionally, technology development has typically revolved around the functionality, usability, efficiency and reliability of technologies.
- However, AI technology needs a broader discussion on its societal acceptability.
- It impacts on moral (and political) considerations.
- It shapes individuals, societies and their environments in a way that has ethical implications.

- The interpretation of ethically relevant concepts can change with technologies (consider what “privacy” meant before social media).
- Furthermore, when new technologies are introduced, users often apply them for purposes other than those originally intended.
- This reforms the ethical landscape, and forces us to reflect and analyze the ethical basis of technology over and over again.

Ethical Frameworks

- Ethical frameworks are attempts to build consensus around values and norms that can be adopted by a community – whether that's a group of individuals, citizens, governments, businesses within the data sector or other stakeholders.
- Various organisations have participated in developing an ethical framework for AI.
- Naturally, their views differ in some respects, but there's also been an emerging consensus to them.
- According to a recent study (Jobin et al 2019), AI ethics has quite rapidly converged on a set of five principles.

Principles of AI Ethics

- non-maleficence
- responsibility or accountability
- transparency and explainability
- justice and fairness
- respect for various human rights, such as privacy and security

- Should we use AI for good and not for causing harm? (the principle of beneficence/ non-maleficence)
- Who should be blamed when AI causes harm? (the principle of accountability)
- Should we understand what, and why AI does whatever it does? (the principle of transparency)
- Should AI be fair or non-discriminative? (the principle of fairness)
- Should AI respect and promote human rights? (the principle of respecting basic human rights)

- The rest of this course will focus on these principles of AI ethics.
- We will analyze what these concepts imply and how they can be interpreted, in the fashion of traditional philosophy: concept analysis.
- We will also look at how these concepts are being applied in practice, discuss their problems and mention some open questions regarding these principles.
- We will look at the project of AI ethics as a whole. We will be asking the “cui bono” question: who is AI ethics for, and who or what is left out?

- Lastly, we want to note that when speaking of AI and the social implications, AI ethics is the first on the list.
- But there are other theoretical frames for looking at ethical codes for algorithmic, data-driven systems.
- For example, questions of the social implications of AI come up in fields like algorithmic cultures, gender studies and media studies, amongst numerous others.
- Correspondingly, the cognitive and psychological aspects of human-machine interaction shapes the question of appropriate ethical framework for AI.
- Simply, there is a lot more to AI ethics than just data or algorithm ethics.

Thank You