# AI and Ethics

Lecture 4

# What should be done

- **The principle of beneficence**
- *“AI inevitably becomes entangled in the ethical and political dimensions of vocations and practices in which it is embedded. AI Ethics is effectively a microcosm of the political and ethical challenges faced in society.” --Brent Mittelstadt*
- The principle of **beneficence** says “do good”,
- The principle of **non-maleficence** states “do no harm”.
- Although these two principles may look similar, they represent distinct principles.

- Beneficence encourages the creation of beneficial AI (“AI should be developed for the common good and the benefit of humanity”),
- while non-maleficence concerns the negative consequences and risks of AI.

- AI ethics have been primarily concerned with the principle of non-maleficence.
- Discussion has focused mostly on questions of how developers, manufacturers, authorities, or other stakeholders should minimize the ethical risks such as
  - – discrimination, privacy protection, physical and social harms – that can arise from AI applications.
- Often, these discussions are stated in terms of intentional misuse, malicious hacking, technical measures, or risk-management strategies.

- Critics claim that the emphasis on non-maleficence makes ethics a matter of finding technical solutions for technical problems.
- Moral problems are seen as things that can be solved by technical “fixes”, or by good design alone.
- The wider ethical and societal context in which technical systems are embedded is forgotten.
- Many significant issues that direct the control, governance and societal dimensions of AI are ignored.

- Technology researcher Evgeny Morozov calls this “tech solutionism” – the conviction that problems caused by technology can always be fixed by more technology.
- As a result, deep and difficult ethical problems are oversimplified and unanswered.
- One of the questions is the problem of the “common good”.
- What, exactly, does that mean?
- How AI Can be useful for common good?

# The common good – calculating consequences

- Suppose you are the Chief Digital Officer in Health Dept.
- You are asked to consider whether the city's health care organisation should move from "reactive" healthcare to "preventive" healthcare.
- You read a report. It tells about novel, sophisticated machine learning systems that would help health authorities to forecast the possible health risks of citizens.

- These methods produce predictions by combining and analyzing various sources of medical and health care systems.
- By analyzing a large number of criteria data, high-risk individuals could be identified and prioritized.
- These high-risk individuals could proactively be invited to a doctor’s appointment to get proper treatment.

# The benefits

- The report mentions many advantages.
- For example, sickness prevention has a lot of potential to improve the health and quality of life for citizens.
- It would allow better impact estimation and planning of basic healthcare services.
- Preventive healthcare also has the potential to significantly reduce social and healthcare costs.
- These savings, the report emphasizes, could be used for the common good.

# The potential problems

- The report also includes some concerns.
- For eg, the systems raise a number of legal and ethical issues regarding privacy, security, and the use of data.
- The report asks, for example, where is the border between acceptable prevention and non-acceptable intrusion?
- Does the Health Dept./Govt have a right to use private, sensitive medical data for identifying high-risk patients?
- How is consent to be given, and what will happen to people who don’t give their consent?
- What about those people who do not give consent because they are not able to?

- The report also raises the fundamental question of the government’s role:
- if the Govt. has information about a potential health risk and does not act upon the data, is the city guilty of negligence?
- Are citizens treated equally in the physical and digital worlds?
- If a person passes out in real life, we call an ambulance without having explicit permission to do so.
- In the digital world, privacy concerns may prevent us from contacting citizens.

- What do you think about the above example?
- As a Chief Digital Officer, would you promote the use of preventive methods?
- If your answer is something like “yes, the Dept./Govt should seek an ethically and legally acceptable way to use those methods –
- there are so many advantages compared to the possible risks”, you were probably using a form of moral reasoning called "utilitarianism".

# Utilitarianism

- **Utilitarianism** is a family of ethical theories.
- It conceives “benefits” as actions that maximize well-being across all affected individuals.
- Utilitarianism is a version of consequentialism, which states that the consequences of any action are the only standards of right and wrong.
- According to utilitarianists, morally right actions are the ones that produce the greatest balance of benefits over harm for everyone affected.

- Unlike other, more individualistic forms of consequentialism (such as egoism) or unevenly weighted consequentialism (such as prioritarianism), utilitarianism considers the interests of all humans equally.
- However, utilitarianists disagree on many specific questions, such as whether actions should be chosen based on their likely results (act utilitarianism), or
- whether agents should conform to rules that maximize utility (rule utilitarianism).
- There is also disagreement as to whether total (total utilitarianism), average (average utilitarianism) or minimum utility should be maximized.

- For utilitarianists, utility – or benefit – is defined in terms of well-being or happiness.
- For instance, Jeremy Bentham, the father of utilitarianism, characterized utility as "that property… (that) tends to produce benefit, advantage, pleasure, good, or happiness…(or)
- to prevent the happening of mischief, pain, evil, or unhappiness to the party whose interest is considered."
- Utilitarianism offers a relatively simple method for deciding, whether an action is morally right or not.

# Steps to be Taken

- Firstly, we identify the various actions that we could perform
- Secondly, we estimate the benefits and harm that would result from each action
- Thirdly, we choose the action that provides the greatest benefits after the costs have been taken into account

- Utilitarianism provides many interesting ideas and concepts.
- For example, the principle of “diminishing marginal utility” is useful for many purposes.
- According to this principle, the utility of an item decreases as the supply of units increases and vice versa.
- For eg., when you start to work out, at first you benefit greatly and your results get dramatically better.
- But the longer you continue working out, each individual training session has a smaller impact.
- If you work out too often, the utility diminishes and you’ll start to suffer from the symptoms of overtraining.

- Eg.2,  if you eat one sweets, you’ll get a lot of pleasure.
- But if you eat too much sweets, you may gain weight and increase your risk to all kinds of sicknesses.
- This paradox of benefits should always be remembered when we evaluate the consequences of actions.
- What is the common good now may not be the common good in the future.

# The problems of utilitarianism

- Utilitarianism is not a perfect account on moral decision making.
- It has been criticized on many grounds.
- For eg, utilitarian calculation requires that we assign values to the benefits and harm resulting from our actions and compare them with the consequences that might result from other actions.
- But it’s often difficult, to measure and compare the values of all relevant benefits and costs in advance.

- "Risk" is commonly used to mean a likelihood of a danger or a hazard that arises unpredictably, or in a more technical sense, the probability of some resulting degree of harm.
- In AI ethics, harm and risks are taken to arise from design, inappropriate application, or intentional misuse of technology.
- Typical examples are risks such as discrimination, violation of privacy, security issues, cyberwarfare, or malicious hacking.
- In practice, it is difficult to compare the risks and benefits

# Reasons for Not able to Compare Risks and Benefits

- **R**isks and benefits are influenced by value commitments, subjective and diverse preferences, practical circumstances, and personal and cultural factors.
- **H**arm and benefits are not static.
- The marginal utility of an item diminishes in a way that can be difficult to foresee. Moreover, a specific harm or a specific benefit may have different utility value in different circumstances.
- For eg, whether or not the faster car will be more beneficial depends on the intended use of it –
  - if it is intended to be a school bus, then we should prioritize safety, but if it is used as a racing car, then the answer may be different.

- **R**eal-world situations are typically so complex that it is difficult to foresee or compare all the risks and benefits in advance.
- For eg, let’s analyze the possible consequences of military robotics.
- Although contemporary military robots are largely remotely operated or semi-autonomous, over time they are likely to become fully autonomous.
- According to some estimates, robots reduce civilian and military casualties.
- But according to other estimates, they do not reduce the risk to civilians.
- Statistically, in the first decades of war in the 21st century, robotic weaponry has been involved in numerous killings of both soldiers and noncombatants.
- The possibility to use various techniques – such as adversarial patches (which interrupt a machine’s ability to properly classify images) – to fool and manipulate automated weapons complicates the situation by increasing the specific risk of causing harm to civilians.
- The overall level of risks is also dependent on the ease in which wars might be declared if robots are taking most of the physical risk.

- **U**tilitarianism fails to take into account other moral aspects.
- It is easy to imagine situations where developed technology would produce great benefits for societies, but its use would still raise important ethical questions.
- For eg., let’s think about the case of a preventive healthcare system.
- The system may indeed be beneficial for many, but it still forces us to ask whether fundamental human rights, such as privacy, matter. Or
- What happens to the citizen’s right not to know about possible health problems?
- Many of us would want to know if we are in a high-risk group, but what if someone does not want to know? Can a Govt. force that knowledge on them? Or,
- How can we make it sure that everyone has equal access to the possible benefits of a preventative system?

# Nozick’s Utility Monster

- One of the biggest difficulties with utilitarianism is the question of utility: what is it really?
- Technically, utility is only a measure (a numeric quantity) that describes some kind of underlying “good” which we want to maximize.
- Say, pleasure, or well-being (which hedonist philosophers would claim to be the same thing).
- Pleasure is at least to some extent a subjective experience, and utility, as a measure, should transform it into an intersubjectively comparable number.
- That is a high bar to reach.

- Assuming such a measure as utility does in fact exist, philosopher Robert Nozick presents the following puzzle.
- There is a creature called the Utility Monster. Their hedonistic mind is wired so that, given any resource, they will receive more pleasure from it than any other individual would.
- They simply enjoy apples, cars, coffee, freedom, etc., more than anybody else does.
- This means that they gain more utility from them, and if we are morally obligated to maximize the utility produced by the resources we have, the conclusion is clear: everything we have to the Utility Monster. Nothing to anybody else.
- Does this make utilitarianism unpalatable?
- Is there a way for the utilitarian to argue that the puzzle Nozick posed is not really a problem?

# Common good and well-being

- Despite the problems outlined earlier, the principles of utilitarianism may help us to consider the immediate and the less immediate consequences of our actions.
- One should remember that in real life, defining “common good” requires a diversity of viewpoints.

# What is "well-being"?

- Often, the term “common good” is taken to be synonymous with “well-being”. But what is well-being?
- The roots of well-being research are in ancient Greece, where philosophers such as Aristotle focused on how to achieve “the good life”.
- Since then, the search for the good life has been a constant topic handled by different disciplines.
- Today, research on fields such as in psychology, economics, and social sciences addresses well-being in terms of "the biological, personal, relational, institutional, cultural, and global dimensions of life".
- These dimensions cover factors such as physical and mental vitality, social satisfaction, and a sense of personal achievement and fulfillment.

# Theories of Well-being

**The subjective theories.**

- These focus on questions such as
  - how people feel as they go about their daily lives, or
  - how a person evaluates their lives.
- This type of psychological well-being is often described as
  - the experience of high life satisfaction,
  - high levels of pleasant emotions and
  - moods, and low levels of negative emotions and moods.

**The eudaimonic theories.**

- These consider well-being primarily as the outcome of positive goal pursuits.
- The eudaimonic perspective differentiates well-being from the satisfaction of desire.
- Well-being and subjective happiness should not be equated because the pleasure-producing outcomes that underlie subjective happiness do not necessarily promote wellness and well-being.
- Instead, well-being can be taken to require components such as autonomy, environmental mastery, personal growth, positive relations with others, a sense of having a purpose in life, and self-acceptance.
- These dimensions describe well-being as an overall positive evaluation of oneself, acceptance of one’s past life and individual talents as a member of a community, the belief that one’s life is meaningful, and a sense of self-determination.

**The social theories.**

- In these, well-being is approached in terms of social factors, such as
  - integration, contributions to social life, social coherence, and social acceptance.
- Well-being is dependent on the degree to which an individual is functioning well in their social environments.

- Surveys like The World Happiness Report provide examples of this holistic approach to well-being.
- The report is an annual publication of the United Nations Sustainable Development Solutions Network.
- It contains articles and rankings of national happiness based on respondent ratings of their own lives, which the report also correlates with various life factors.
- (As of March 2020, Finland was ranked the happiest country in the world three times in a row.)

- Moreover, researchers develop constantly novel ways to approach well-being.
- For eg, big data is nowadays utilized for well-being research in many ways.
- Contemporary methods include the more advanced analysis of demographic and socio-economic data,
- For eg. utilization of text mining tools in any written documents – such as Twitter feeds, Facebook posts, or other social media data, as well as the analysis of digital footprints and even facial features.

# Coomon Good Approach for AI

- The common good approach requires that everyone should have access to the benefits of AI.
- This highlights the importance of ensuring that potential benefits of AI do not accumulate unequally, and are made accessible to as many people as possible.
- AI should be aligned with values, goals, and norms, respecting cultural and individual diversity to a sufficient degree.

- The common good is not a singular, but a plural.
- Identifying social and moral norms of the specific community in which an AI will be deployed is, thus, obligatory.
- It is the only way to bring AI’s potentially significant and diverse benefits to society and facilitate, among other things, greater well-being and welfare for all.

# Thank you