5CS037-Concepts and Technologies of AI
Tutorial-06

The Ethics of Artificial Intelligence.

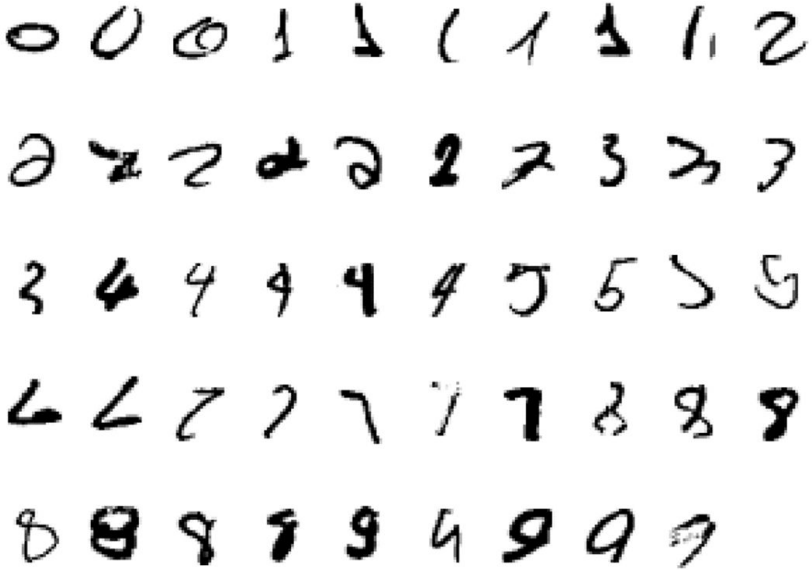
Siman Giri

Lecture Revision

Things to remember from lectures.

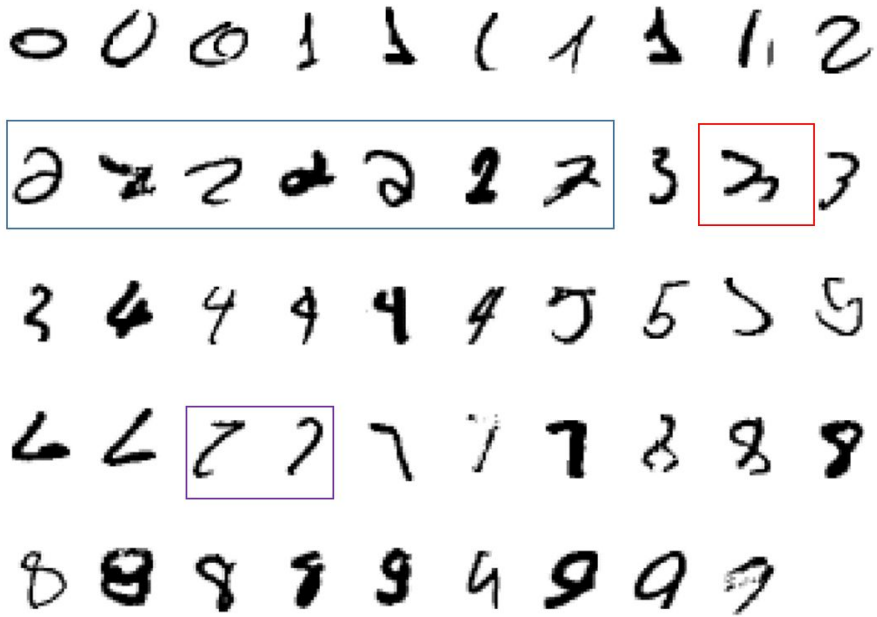
Why do we need Learning?

- Identify 2 from following set of images:
  - It is very hard to say what makes 2.
  - What distinguishes 2 from 7?



Why do we need Learning?

- Identify 2 from following set of images:
  - It is very hard to say what makes 2.
  - What distinguishes 2 from 7?



What is Machine Learning?

- Some popular definition from legends of the field:
  - "Learning is any process by which a system improves performance from experience". **-- Herbert Simon**
  - Definition by **Tom Mitchel(1998):**
    - Machine Learning is the study of algorithms that:
      - **Improve their performance P**
      - **At some task T**
      - **With experience E**
    - A well defined learning task is given by **<P, T, E>.**
  - "Field of study that gives computers the ability to learn without being explicitly programmed." **- - Arthur Samuel ,1959 (an AI pioneer at IBM).**

Tasks that can use Machine Learning.

- Handwriting recognition learning problem
  - Task T: Recognizing and classifying handwritten words within images
  - Performance P: Percent of words correctly classified
  - Training experience E: A dataset of handwritten words with given classifications
- A robot driving learning problem
  - Task T: Driving on highways using vision sensors
  - Performance measure P: Average distance traveled before an error
  - Training experience E: A sequence of images and steering commands recorded while observing a human driver
- A chess learning problem
  - Task T: Playing chess
  - Performance measure P: Percent of games won against opponents
  - Training experience E: Playing practice games against itself

Machine Learning: Types.

- There are many different problem classes in machine learning.
- They vary according to **what kind of data** is provided and **what kind of conclusions** are to be drawn from it.
- Some-popular kind are:
  - **Supervised Learning**
  - Unsupervised Learning
  - Reinforcement Learning
- In this course, we will focus on **classification and regression** (two examples of supervised learning) and will touch on unsupervised learning if time permits.

Components of Machine Learning: Dataset

Labeled Dataset

Features

Label

date	lat	long	temp	humidity	cloud_coverage	wind_direction	atmp_pressure	rainfall
2021-09-09	49.71N	82.16W	74	20	3	N	18.6	.01
2021-09-09	32.71N	117.16W	82	42	6	SW	29.94	.23

Example

Features

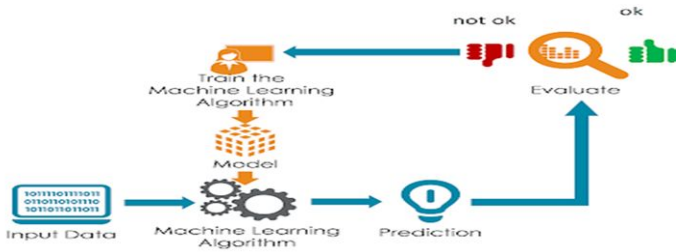
date	lat	long	temp	humidity	cloud_coverage	wind_direction	atmp_pressure
2021-09-09	49.71N	82.16W	74	20	3	N	18.6
2021-09-09	32.71N	117.16W	82	42	6	SW	29.94

Example

Unlabeled Dataset

Elements of Machine Learning

- There may be hundreds of machine learning algorithms, all of those algorithms must have following three attributes:
  - **A Decision Process (Representation/Model):**
    - Machine learning algorithms(Models) are used to make inference or estimate of an output based on input data – labeled or unlabeled.
  - **An Error Function (Evaluation):**
    - A performance metric used to evaluate the estimate of a model.
    - Metrics depends on types of learning (supervised or unsupervised) and types of task (Classification or Regression).
  - **An model Optimization Process:**
    - An automated algorithm or process used to update parameters of machine learning models until threshold or accepted evaluation metric has been achieved.



Supervised Machine Learning!!

- **Data:**
  - For Supervised Learning Setup, **training data** comes in pairs of inputs **(x, y):** where $X \in R^d$ is the input instance and Y its label, which can be written as:
    - $D = \{(x_1, y_1) \dots (x_n, y_n)\} \subseteq R^d * C$
    - Where:
      - R^d: d-dimensional feature space.
      - x_i: input vector of the i^{th} sample.
      - y_i: label of the i^{th} sample.
      - C: label space.

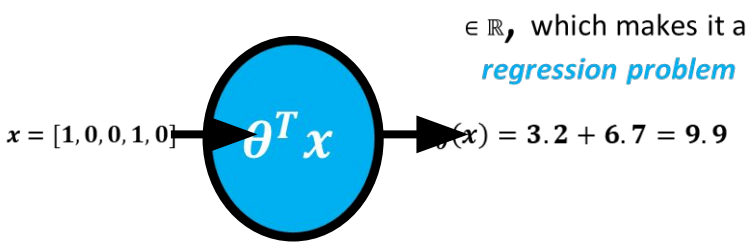
Supervised Machine Learning!!

- **Data: label Space**.
  - There can be multiple scenario for the label space c.

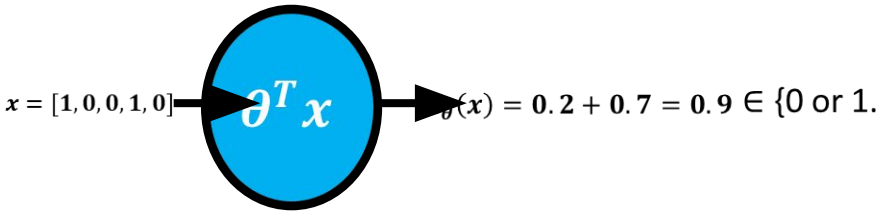
Binary Classification	$c = \{0 \text{ or } 1\}$	E.g.: An email is either spam or not a spam.
Multi Class Classification	$c = \{1,2,..k\}(k \geq 2)$	E.g.: Traffic sign Classification.
Regression	$c = \mathbb{R}$	E.g.: Height of the person.

Regression Vs. Classification.

- Regression:
  - What are the possible outputs of the linear regression function $\boldsymbol{h_\theta(x) = \theta^T x}$?
    - Real-valued Outputs.



- Classification:
  - What are the possible outputs of the linear regression function $\boldsymbol{h_\theta(x) = \theta^T x}$?
    - Discrete Outputs.



Learning Goals

- Discuss and observe Ethical flaws of AI.
- What can be the Ethical issues that may arise in the near future of AI?
- What can be the Fundamental principle of building Ethical AI?

Ethics of AI:
Example Case Studies of Modern Challenges and Major Debates.

1. Cases of AI Enabled Discrimination.

1.1 Gender Bias: Case1

- **Percentage of women in top 100 Google image search results for CEO: 11%**
- **Percentage of U.S. CEOs who are women: 27%**

- **Percentage of women in the top 100 Google image search results for telemarketers: 64%**
- **Percentage of U.S. telemarketers who are women: 50%**

1.1 Gender Bias: Case-2

- In 2014, Amazon started to develop and use AI programs to mechanize highly time intensive human resources (HR) work, namely the shortlisting of applicants for jobs.
  - Amazon "literally wanted it to be an engine where I'm going to give you 100 résumés, it will spit out the top five, and we'll hire those" (Reuters2018).
  - The AI tool was trained on CVs submitted over an earlier ten-year period and the related staff appointments.
  - Following this training, the AI tool **discarded the applications of female applicants**, even where no direct references to applicants' gender were provided.
  - Given the **predominance of successful male applicants in the training sample**, Amazon found that the system penalized language such as "women's chess club captain" for not **matching closely** enough the successful **male job applicants** of the past.
  - While developers tried to modify the system to avoid gender bias, Amazon abandoned its use in the recruitment process in 2015 as a company "committed to workplace diversity and equality" (ibid).

1.2 Law Enforcement and Policing: Case1

- Glenn Rodríguez had been arrested at the age of 16 for his role in the armed robbery of a car dealership, which left one employee dead.
  - In 2016, 25 years later, he applied to the parole board of the Eastern Correctional Facility in upstate New York for early release. **He had a model rehabilitation record at the time (Wexler2017b).**
    - **Parole was denied.** The justification given by the board was that an **AI system called COMPAS** had **predicted him** to be **"high risk"** and the board "**concluded that … release to supervision is not compatible with the welfare of society**" (Wexler 2017a).
  - The parole board had **no knowledge** of how the **COMPAS risk score was calculated**, as the company that had developed the system considered their **algorithm a trade secret** (ibid).
  - Through cross-referencing with other inmates' scores, Rodríguez found out that the reason for his high-risk score was a subjective personal view given by prison guards, who may have been influenced by racial prejudices. In the end, he was released early.
  - However, "had he been able to examine and contest the logic of the COMPAS system to prove that its score gave a distorted picture of his life, he might have gone home much earlier" (Wexler 2017b)

1.3 Bias based on Ethnicity: Case1

- In 2016, a 22-year-old engineering student from New Zealand had his passport photo rejected by the systems of the New Zealand department of internal affairs because his eyes were allegedly closed.
  - The student was of **Asian descent and his eyes** were open.
  - The automatic photo recognition tool declared the photo invalid and the student could not renew his passport.
    - He later told the press very graciously: **"No hard feelings on my part, I've always had very small eyes and facial recognition technology is relatively new and unsophisticated" (Reuters 2016).**
- Similar cases of **ethnicity-based errors** by passport photo recognition tools have affected dark-skinned women in the UK.
  - **"Photos of women with the darkest skin were four times more likely to be graded poor quality, than women with the lightest skin" (Ahmed 2020).**
    - **For instance, a black student's photo was declared unsuitable as her mouth was allegedly open, which it in fact was not (ibid).**

1.4 Ethical Questions: AI-Enabled Discrimination.

- The reproduction of **biases and resulting discrimination** are among the **most prominent ethical concerns about AI** (Veale and Binns 2017; Access Now Policy Team 2018).
- **Bias** has been described as the "**one of the biggest risks associated with AI**"(PwC2019: 13).
  - AI systems are only **as good as the data they're trained on** and **the humans that build them**.
    - If a résumé-screening **machine-learning tool is trained on historical data**, such as résumés collected from a company's previously hired candidates, **the system** will **inherit** both the **conscious and unconscious preferences** of the **hiring managers** who made those selections (Heilweil2019).
  - **Deep neural networks** for **image classification** … are often trained on **ImageNet** … More than **45% of ImageNet data**, which fuels research in computer vision, comes from the **United States**, **home to only 4% of the world's population**.

Ethics of AI:
Example Case Studies of Modern Challenges and Major Debates.

2. Cases of Surveillance and Privacy Violation.

2.1 Social Credit Scoring: Case1

- China is one of the world's leading nations in AI development. It embraces the use of large amounts of data that it collects on its citizens, for instance in its **social credit scoring system**.
  - This system uses a large number of data points, including social media data, local government data and citizens' activities, to calculate a trustworthiness score for every citizen.
  - Several data platforms are used to integrate data into "a state surveillance infrastructure" (Liang et al. 2018).
  - **High scores** lead to the **allocation of benefits**, such as lower utility rates and favorable booking conditions, whereas low scores can **lead to the withdrawal of services** (Raso et al.2018).
  - Within China, the system benefits from **high levels of approval** because Chinese citizens "interpret it through frames of benefit-generation and promoting honest dealings in society and the economy instead of privacy-violation" (Kostka2019: 1565).

2.2 Biometric Surveillance: Case 2

- "Nijeer Parks is the third person known to be arrested for a crime he did not commit based on a bad face recognition match" (Hill 2020).
  - Parks was **falsely accused of stealing and trying to hit a policy officer** with his car based on **facial recognition software – but he was 30 miles away at the time**.
  - "**Facial recognition** ... [is] **very good with white men**, **very poor on Black women** and not so **great on white women**, even"(Balli 2021).
  - It becomes particularly problematic when "**the police trust the facial recognition technology more than the individual" (ibid).**

2.4 Ethical Questions: Privacy and Surveillance.

- The above two case studies show that the **analysis of personal data through AI systems** can lead to **significant harms**.
  - AI is by far not the only threat to privacy, but it **adds new capabilities** that can either **exacerbate existing threats**,
    - for example by **automating mass surveillance based on biometric data**,
  - or add **new angles to privacy concerns**,
    - for example by **exposing new types of data**, such as **genetic data**, to the possibility of **privacy violations.**
- Finally, like most other fundamental rights, **privacy is not an absolute right**.
  - Personal privacy **finds its limits when it conflicts with other basic rights or obligations**, for example **when the state compiles data in order to collect taxes or prevent the spread of diseases.**
- The balancing of privacy against other rights and obligations therefore plays an important role in finding appropriate mitigations for privacy threats.

Ethics of AI:
Example Case Studies of Modern Challenges and Major Debates.

3. Cases of Manipulation.

3.1 Election Manipulation: Case 1

- The 2008 US presidential election has been described as the first that "relied on large-scale analysis of social media data, which was used to improve fundraising efforts and to coordinate volunteers" (Polonski2017).
- The increasing availability of large data sets and AI-enabled algorithms led to the recognition of new possibilities of technology use in elections.
  - In the early 2010s, **Cambridge Analytica**, a voter-profiling company, wanted to become active in the 2014 US midterm election (Rosenberg et al.2018).
  - The company attracted a **$15 million investment from Robert Mercer**, a **Republican donor**, and engaged Stephen Bannon, who later played a key role in President Trump's 2016 campaign and was an important early member of the Trump cabinet.
  - Cambridge Analytica **lacked the data required for voter profiling**, so it solved this problem with **Facebook data (Cadwalladr and Graham-Harrison2018).** Using a permission to **harvest data** for **academic research purposes that Facebook had granted to Aleksandra Kogan**, a researcher with links to **Cambridge University**, the company harvested not just the data of people who had been **paid** to take a personality quiz, but **also that of their friends**. This allowed Cambridge Analytica to harvest in total **50 million Facebook profiles**, which **allowed the delivery of personalized messages** to the profile holders and also – importantly – a **wider analysis of voter behavior**.

3.2 Pushing Sales: Case 2

- Human beings do not feel and behave the same way all of the time; they have ups and downs, times when they feel more resilient and times when they feel less so.
- A 2013 marketing study suggests that one can identify typical times when people feel more vulnerable than usual.
  - US women across different demographic categories, for example, have been found to **feel least attractive on Mondays**, and **therefore possibly more open to buying beauty products** (PHD Media2013).
    - This study goes on to suggest that such insights can be used to develop bespoke marketing strategies.
    - While the original study couches this approach in positive terms such as "encourage" and "empower", independent observers have suggested that it may be the "grossest advertising strategy of all time" (Rosen 2013).

3.3 Ethical Questions: Manipulation.

- There are numerous interventions which claim that AI can influence human behavior (Whittle 2021), for example by understanding cognitive biases and using them to further one's own ends (Maynard 2019).
- In particular the collecting of data from social media seems to provide a plausible basis for this claim, where manipulation (Mind Matters2018) is used to increase corporate profits (Yearsley2017).
- However, any such interventions look different from other threats to our freedom to act or to decide, such as incarceration and brainwashing.
- Another answer to the question why AI-enabled manipulation is ethically problematic
- is that it is based on privacy infringements and constitutes surveillance.
- Facebook users in the Cambridge Analytica case were not forced to vote in a particular way but received input that influenced their voting behavior.
    - AI can have (and likely already has) an adverse impact on democracy, in particular where it comes to:
        - (i) **social and political discourse, access to information and voter influence**,
        - (ii) **inequality and segregation and**
        - (iii**) systemic failure or disruption. (Muller 2020: 12)**

Ethics of AI:
Ethical Dilemmas and Accountability.

4. Example Cases.

4.1 Ethical Dilemmas: Autonomous Car

- An autonomous car is a vehicle that is capable of sensing its environment and moving with little or no human involvement.
  - to move safely and to understand its driving environment, an enormous amount of data needs to be captured and processed
- Imagine an autonomous car with broken brakes going at full speed towards a grand-mother and a child. By deviating a little, one can be saved.
- Who would you choose, the grandmother or the child? Do you think there is only one right answer?

4.1 Ethical Dilemmas: Autonomous Car

- Example:
  - If the machine with intelligence act, will they themselves be responsible, liable, or accountable for their actions? Or should the distribution of risk perhaps take precedence over discussions of responsibility?

Self-driving cars

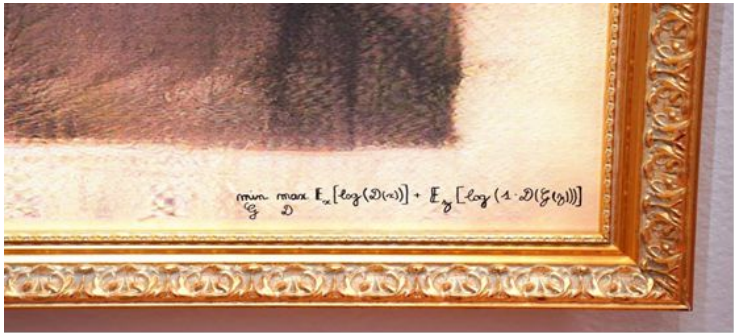
Tesla car that crashed and killed driver was running on Autopilot, firm says

- Company says driver took no action despite system's warnings
- Uber settles with family of woman killed by self-driving car

4.2 Ethical Dilemmas: Ownerships

- Example: AI creates art.
- In 2019, Huawei announced that an AI algorithm has been able to complete the last two **movements of Symphony No.8**, the unfinished composition that **Franz Schubert** started in 1822, 197 years before.
- Researchers try to create AI tools that can **generate music , paintings and poetry**, that's just as compelling as the human-made kind.
- ***Pierre Fautrel*** *created the artwork, which sold at Christies for $430,000 (£335,000) in October 2018*

4.2 Ethical Dilemmas: Ownerships



- ***Pierre Fautrel*** *created the artwork, which sold at Christies for $430,000 (£335,000) in October 2018*
- So what happens when AI has the capacity to create works of art itself?
- If a human author is replaced by machines and algorithms, to what extent copyrights can be attributed at all?
- Can and should an algorithm be recognized as an author, and enjoy the same rights as an artist?

4.3 Ethical Dilemmas: Employment.

- Classic automation replaced human muscle, whereas digital automation replaces human thought or information-processing
- According to the World Economic Forum's "The Future of Jobs Report 2020", AI is expected to replace 85 million jobs worldwide by 2025.The report goes on to say that it will also create 97 million new jobs in that same timeframe.
- Will we be ready for the transition?

4.4 Ethical Dilemmas: LLMs(chatGpt)

- LLMs: Large Language Models.
- The study by (Sebastian Porsdam and Brian D. Earp, Oxford University) reveals that LLMs like ChatGPT pose crucial questions regarding the **attribution of credit and rights for useful text generation**, **diverging from traditional AI responsibility debates that primarily focused on harmful consequences**.
- A key finding of the research, according to co-authors Sven Nyholm and John Danaher, '**is that while human users of these technologies cannot fully take credit for positive results generated by an LLM, it still seems appropriate to hold them responsible for harmful uses, such as generating misinformation, or being careless in checking the accuracy**' of generated text.
- The paper points out that LLMs may be helpful in education, but warns that they are error-prone, and overuse might affect critical thinking skills.
  - Education and publishing are particularly in need of rapid action on guidelines for LLM use and responsibility.
  - We should consider adapting assessment styles, rethinking pedagogy, and updating academic misconduct guidance to handle LLM usage effectively.

What Should we do?

- The ethical challenges presented by artificial intelligence (AI) are one of the biggest topics of the twenty-first century.
  - The potential benefits of AI are said to be numerous, ranging from operational improvements, such as the reduction of human error (e.g.in medical diagnosis), to the use of robots in hazardous situations (e.g. to secure a nuclear plant after an accident).
  - At the same time, AI raises many ethical concerns, ranging from algorithmic bias and the digital divide to serious health and safety concerns.
- AI system designed or build must be able to take into account **societal values, moral and ethical considerations**; weigh the respective **priorities of values held by different stakeholders in various multicultural contexts**; **explain its reasoning; and guarantee transparency**. (Dignum 2018: 1, 2)
-

5. What is Ethics?

Introduction to Ethics.

5.1 Defining Ethics.

- Ethics is a set of moral principles which help us discern between right and wrong.
- ethics as the principles that guide our behavior toward making the best choices that contribute to the common good of all. Ethics is what guides us to tell the truth, keep our promises, or help someone in need

5.2 Ethics and AI.

- Ethical AI is artificial intelligence that adheres to well-defined ethical guidelines regarding fundamental values, including such things as individual rights, privacy, non-discrimination, and non-manipulation.
- AI Ethics is a set of guidelines that advise on the design and outcomes of artificial intelligence.
- There are not any known universal ethical guidelines,
- Recommendation are made by developing agencies, big corporates themselves, group of researchers and academician.

5.3 Field of Ethics and AI.

- The field of AI ethics has boomed into a global enterprise with a wide variety of players. Yet the ethics of artificial intelligence (AI) is nothing new.
- The concept of AI is almost 70 years old (McCarthy et al.2006) and ethical concerns about AI have been raised since the middle of the twentieth century (Wiener1954; Dreyfus 1972;Weizenbaum1977).
- The debate has now gained tremendous speed thanks to wider concerns about the use and impact of better algorithms, the growing availability of computing resources and the increasing amounts of data that can be used for analysis(Hall and Pesenti2017).
- With new uses of AI, AI ethics has flourished well beyond academia.

AI Ethicist:

[I]t is not the role nor to be expected of an AI Ethicist to be able to program the systems themselves. Instead, a strong understanding of aspects such as the difference between supervised and unsupervised learning, what it means to label a dataset, how consent of the user is obtained – essentially, how a system is designed, developed, and deployed – is necessary. In other words, an AI Ethicist must comprehend enough to be able to apprehend the instances in which key ethical questions must be answered (Gambelin 2021).

6. Building Ethical AI.

6.1 Generic Principles for the development, implementation and use of AI

- **Human rights**: AI should be developed and implemented in accordance with international human rights standards.
- **Inclusiveness**: AI should be inclusive, aiming to avoid bias and allowing for diversity and avoiding a new digital divide.
- **Flourishing**: AI should be developed to enhance the quality of life.
- **Autonomy**: AI should respect human autonomy by requiring human control at all times.
- **Explainability**: AI should be explainable, able to provide insight into its functioning.

6.1 Generic Principles for the development, implementation and use of AI

- **Transparency**: The data used to train AI systems should be transparent.
- **Awareness and literacy**: Algorithm awareness and a basic understanding of the workings of AI are needed to empower citizens.
- **Responsibility**: Developers and companies should take into consideration ethics when developing autonomous intelligent system.
- **Accountability**: Arrangements should be developed that will make possible to attribute accountability for AI-driven decisions and the behavior of AI systems.
- **Democracy**: AI should be developed, implemented and used in line with democratic principles.
- **Good governance**: Governments should provide regular reports about their use of AI in policing, intelligence, and security.
- **Sustainability**: For all AI applications, the potential benefits need to be balanced against the environmental impact of the entire AI and IT production cycle

Useful Reference for Term Paper

- Generative AI entails a credit–blame asymmetry (https://www.nature.com/articles/s42256-023-00653-1)
- Practical and ethical challenges of large language models in education: A systematic scoping review (https://bera-journals.onlinelibrary.wiley.com/doi/full/10.1111/bjet.13370)
- PRELIMINARY STUDY ON THE ETHICS OF ARTIFICIAL INTELLIGENCE by UNESCO Working Group (https://unesdoc.unesco.org/ark:/48223/pf0000367823)
- Set of fictional case studies that are designed to prompt reflection and discussion about issues at the intersection of AI and Ethics. (https://aiethics.princeton.edu/case-studies/case-study-pdfs)
- Ethics of Artificial Intelligence and Robotics. (https://plato.stanford.edu/entries/ethics-ai)
- The Ethics of Artificial Intelligence (https://intelligence.org/files/EthicsofAI.pdf)

Thank You any Question!!!