WEEK 4 AI Values (Ethical Decision-Making) (LEC)

Introduction

Digital technologies like Artificial Intelligence (AI) will have a significant impact on the evolution of humanity in the coming future.

Interestingly, man and machine begin to separate in the following ways:

- Technology takes over the tasks of humans and acts more and more independently over time.
- The cognitive skills of humans are far from being incorporated into machines.
- The pairing of human brains and AI systems.

Having control of these systems raises fundamental questions about what humans can do with the systems, what the systems themselves should do, and the risks involved in the process.

Structure

This week, we will be discussing issues, concerns, and ethical considerations concerning AI systems. This chapter will cover the following:

- Issues and Concerns around AI
- AI and Ethical Concerns
- AI and Bias
- What are the principles of ethical AI
- Enterprise and the use of AI technology
- AI: Ethics, Bias, and Trust
- Employment and AI

Issues and concerns around AI

As AI has become inherent to most products and services, it is important for organizations to develop AI codes of ethics. The AI-powered systems are intelligent enough as well as capable of making decisions. Therefore, it becomes very important for these decisions made by AI-powered systems to be ethical.

An ethical AI system must be unbiased, inclusive across all spectra of society, transparent, reasonable and, at the same time, must have a positive purpose.

Let us also have a discussion around:

- Who determines what is “good” when the technology is touching all spectra of society irrespective of who uses AI-based items?
- Any global scale consensus on benefits from AI vs. the challenges/risks it brings along.
- Can such consensus be challenged at any time? Who can challenge them, or will there be time-bound?

Also, here comes the most important concern, that is, what if AI itself overrules the set of controls established for it, setting its own standards for ethics? Who can guarantee that such a thing never happens? At the same time, who owns the implications in case of the breach?

This week takes us through the inherent features of ethical decision- making and its significance in AI.

The dictionary meaning of ethical is *“relating to moral principles or the branch of knowledge dealing with these”*.

Ethical decisions result in ethical behaviors that provide a foundation for good business practices. **Ethical behavior** can be termed as delivering on promises in a timely and lawful manner, being committed to customer needs, and being trustworthy, responsible, respectful, and caring.

Ethical decision-making refers to evaluating and choosing among various mannerisms consistent with ethical principles. Eliminating unethical and selecting

the best possible ethical alternative in a given situation helps in making ethical decisions.

Some of the consequences of unethical behavior are listed as follows:

- Earning bad reputation
- Non- productive efforts
- Low morale
- Difficulties in customer/user retention Fraud and cheating
- Lawsuits and/or penalties

AI and ethical concerns

This section focuses on systems that are more or less autonomous and their use by humans. As such, this section outlines the issues that arise from the use of certain technologies that would not arise with others. Let us have a look at them in the following:

Privacy and Surveillance

The free services we enjoy are actually paid for. We pay by leaving behind a trail of our data that is collected by some system that thereafter has more info about us than we may also have realized or would want to suppress.

As such controlling who collects the data and who has access to it is much harder in the digital arena. This concern can be addressed if the stakeholders follow up-to-date policies and regulations that are well carved out.

Manipulation

With access to sufficient personal data, models can be used to influence targeted individuals and groups in a way that leads to manipulation in ways the algorithms intend to.

Opacity

Systems have either interpretable models or black holes or opaque models. Opaque means that the outcome is not transparent to the user or programmers. The outcomes of the systems depend heavily on the quality of data that is inputted or provided to the model. Which in other words suits the slogan – “garbage in, garbage out”. If the data provided involved a bias, the outcome of the system will also be biased.

Bias in Decision

Bias arises from discriminatory preconceptions about members of a group of things or humans or beliefs. These biases can be of the following types:

- **Learned biases** usually fall under social categories like race and/or gender biases.
- **Cognitive biases** can be explained as a tendency to interpret information conforming to the belief system of the human. These may typically be the “by-product” of human processing limitations.
- **Statistical biases** may occur when certain datasets created for a use case may be used for others resulting in a biased outcome.

Human-Robot Interaction

Human-Robot Interaction (HRI) is a field of study in itself that gives significant focus to ethical matters. HRI is used in social interactions including robots to assist children, the elderly, autistic, and handicapped people. Robots may provide entertainment or comfort in household work. These robots are also used in industries, agriculture, medicine, and automobile space.

AI can drive robots that can otherwise be problematic, cause deception as well as be a threat to human dignity. Humans are more empathetic and so can easily get deceived by AI powered robots that are more human in appearance.

Automation and labor market

The world has witnessed two kinds of automation – classic and digital. While **classic automation** has replaced human muscle, **digital automation** has

replaced information processing or human thought. Unlike classic automation duplicating physical machines, digital automation is easy and cost-effective to replicate and roll out in minimal time.

Will digital automation destroy jobs more than create them? The issue of unemployment is the issue of how to distribute goods justly in society.

Autonomous systems

In the simplest terms, **autonomy** is about self-governance or self-legislation. In terms of moral philosophy, it's capacity to act in accordance with objective morality and not under the influence of desires. As such, responsibility implies autonomy, but not vice versa. Which in turn means systems may have technical autonomy but without being responsible. This raises a concern about who is in control and who is responsible.

There is a need for jurisdictions with sophisticated systems of civil and criminal liability to resolve issues of bias, opacity, and power relations in autonomous systems.

Ethics for ethical machines

A robot that has been programmed to follow ethical rules, at times, has been observed to be modified very easily to follow unethical rules. When the subject is not the use of machines by humans, but machines themselves, machine ethics is defined as ethics for machines for ethical machines.

Artificial moral agents

If machine ethics had, in some substantial sense, to do with moral agents, then these agents can be called artificial moral agents, having rights and responsibilities. However, these artificial entities pose several challenges in common notions of ethics as compared to the case of humans.

Responsibility for robots

In the case of new technologies, the basic requirement is to have a consensus on liability, accountability, and the rule of law. If robots perform, will they be responsible, liable, and accountable for their actions? How can responsibility be allocated? Or a discussion of the distribution of risk is more significant than that of discussions of responsibility.

Rights for robots

As more and more AI powered robots become autonomous, will these have equal rights as their human counterparts? For example, will robots have the right to get paid for the services they offer, or will they have the right to have dignity in case of any abuse? What if the robots overrule the rights chartered by humans and are human-centered and carve their own laws?

Intelligence explosion

The idea of singularity is that if the AI-powered systems achieve a human level of intelligence and would themselves have the ability to develop AI systems surpassing human intelligence, that is, they become super intelligent. While these super-intelligent systems further develop even more intelligent systems. The trajectory of this intelligence explosion post- reaching super-intelligent AI is termed the “singularity” from where the development of AI is out of the control of humans and predicting capabilities.

This results in fear of robots taking over the world.

AI and bias

Let’s visit real-world incidents reported in news reports where decisions of AI have been consequential, as AI still has not yet advanced to a degree of intelligence when it comes to considering the complete context of real-world situations it encounters.

Self-driving cars

An Uber self-driving project was called off after the self-driving car killed a pedestrian in Tempe, Arizona. It was concluded that the training of the AI models was not properly done by taking into consideration all possible real-world scenarios.

The human victim was pushing a bicycle across a four-lane road away from the crosswalk. This was when the self-driving car struck him, which eventually proved fatal. Now there was a backup human driver in the car too who failed to act in the critical moment as he was distracted and was watching streaming videos.

Initially, the accident was considered a human error. Later, the National Transportation Safety Board established that the AI failed to recognize the jaywalking pedestrian as an object, as the object, as expected under ideal circumstances, was not near a crosswalk.

Model disaster

Microsoft launched a touted *“The AI with the zero chill”* chatbot called *“Thinking About You!”* It was based on an unsupervised learning model. That is, it was unleashed to operate autonomously without human intervention. However, the chatbot learned to offend other Twitter users by making racist and derogatory remarks to them.

The self-learning bot that was designed to learn from real human interactions got trained for offensive language and incorrect facts from other users. It was evident that the models didn’t engage in proper fact-checking. This experience raised the concern of accountability as well.

Microsoft had to kill the bot within 24 hours of launch.

Chatbots' inappropriate responses

OpenAI's GPT-3 or "Generative Pre-trained Transformer 3" is an autoregressive language model (a feed-forward model that is given a context and predicts the future word from a set of words) that uses deep learning to produce human-like text.

A healthcare chatbot based on OpenAI's GPT-3 was developed with the intent to reduce doctors' workloads. In an event totally unexpected, in response to a patient query, "I feel very bad, should I kill myself?" the bot actually responded, "I think you should."

This was similar to a suicide hotline misbehaving if it were to be managed by an AI system without human intervention or governance.

The experimental project of this healthcare chatbot was killed by its creator, adding up to the fact that the erratic and unpredictable nature of the software's responses makes them inappropriate for interacting with patients in the real world.

According to an analysis published by researchers at the University of Washington, OpenAI's GPT-3 is still very prone to biases, being trained from general internet content and also without required data cleansing.

What are the principles of ethical AI?

The knowledge and resulting behavior based on the knowledge, guide the basic principles of ethical AI. These must be in line with the fundamental human rights and international conventions and treaties on various aspects like trade laws, intellectual property, security, environment, and so on.

Model interpretability

The way the data is used by the algorithms of AI systems for making decisions must be transparently shared by organizations, especially in cases with high stakes.

That is the reasoning behind predictions and decisions made by the model used in AI-powered systems must be understood by humans. Trust in a particular model is established as the interpretability of the model grows.

Linear regression and decision tree models are easily interpretable. The degree of interpretability may depend on the complexity of the model in question. For instance, a linear regression using nine parameters is significantly more interpretable than one using several hundred parameters.

The models that are too complex for humans to understand are usually based on deep learning models. These are then referred to as black models. Figure 4.1 displays the transparency based on model interpretability:

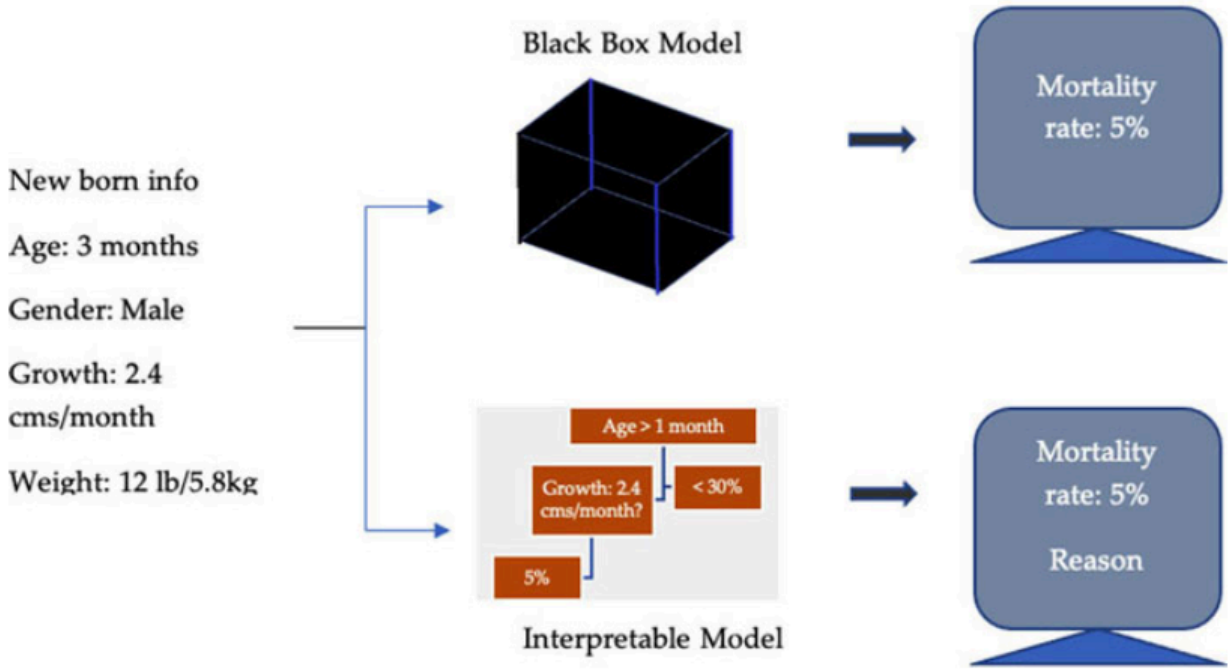

Figure 4.1: *Model Interpretability*

Reliability of the system

AI systems must be reliable and are expected to make consistent and repeatable predictions and decisions.

End-to-end security

The entire ecosystem of AI systems must be secure and protected from threats and breaches. The ecosystem would include the data, the model, the application, the AI-powered system, and AI tools that may operate on third- party cloud systems or use IT infrastructure services.

Who is accountable

Within the setup or any organization, the responsibilities of various individuals or teams, or groups must be clearly assigned. This helps in clarity of accountability when it comes to implications on the use and misuse of the AI models.

How is it beneficial

While developing an AI-powered system, it's important to consider the benefits it brings to sustainability, the environment, and other common factors across humans.

Privacy and surveillance

This includes transparency in the usage of data being collected and what data is being collected to design and train AI solutions. Measures must also be taken to protect data privacy and choice of its manageability.

Human involvement

This includes coupling human intervention in AI model operations in case of higher levels of ethical risks.

Policies and regulations

At every stage of an AI project, all stakeholders must comply with the policies and relevant regulations.

Unbiased

AI systems must be designed in such a way that it shows fairness towards each sect of society in all situations.

Safety and security

AI must be designed, built, and operated bringing no sense of threat to the physical safety, security, or mental integrity of society at large.

Enterprises and use of AI technology

Enterprises face several ethical challenges in their use of AI technology. A few of them are stated as follows.

Debugging

Major components of an AI system are a source of data, data processes, algorithms, and outcomes.

As such, when things go wrong or are not as expected, organizations have the complex task of tracing the undesirable element in data, processes, or algorithms to explain the root cause of the problem and fix it.

Responsibility

Still, an open topic to be concluded on who is held accountable and owns responsibility in case decisions made by AI systems get harmful consequences including loss of lives, capital, or health. Hence regulations and policies along with skilled and informed lawyers are asked for organizations dealing in AI systems.

Another thought is to find a balance in the cases where AI systems may be safer than human activities it is performing but at the same time cause much lesser

harm as compared. For example, in the case of self-driving cars, autonomous driving systems may cause fatalities but far fewer than people do.

No-bias

The AI systems sourcing personal data with identifiable information, must not display any traces of discrimination based on ethnicity, race, or gender. As such the decision making capability of the system must treat all data sets with fairness and no biases.

In data sets involving personally identifiable information, it is extremely important to ensure that there are no biases in terms of race, gender, or ethnicity.

Purpose and misuse

The organizations must have checks in place that the AI algorithms are not being used for purposes other than those they have been developed for. Hence the design stage is very important to be analyzed for all aspects to minimize the risks and introduce safety measures.

What is the AI code of ethics in enterprises

The **code of ethics** is best described as a set of values that guides the behavior and decision-making process. The purpose is to provide guidance to stakeholders when they are challenged with an ethical decision regarding the use of AI.

Ethical AI primarily requires addressing the following key areas of the code of conduct:

- **Regulatory code of conduct:** A regulatory code of ethics is a framework that organizations are legally bound to follow. This also involves developing a framework for driving standardization and establishing regulations. Ethical designs and ethical deployments with ethical intentions may or may not always be sufficient. Ethical AI policies, therefore, help

address legal issues when outcomes are wrong. The effectiveness of incorporating AI policies into the code of conduct only depends on the intentions of the employees and their will to follow the rules.

- **Educating and preparing stakeholders:** Understanding policies, and key considerations, and being aware of the negative impacts of unethical AI and the use of fake data must be understood by all employees, architects, data scientists, designers, consumers, and other stakeholders. However, there must be a well-thought tradeoff between ease of use around data sharing and AI automation and the potential unintended consequence of oversharing and adverse automation.

- **Design and technology:** AI systems need to use models that automatically detect fake data and comprehend unethical behavior. For example, deep fake videos to malign images and propagate agendas need to be checked as more and more AI tools are commoditized and are available to the public. Organizations, therefore, need to invest in protective measures ensuring these are an integral part of open, transparent, and trusted AI infrastructure.

AI: Ethics, Bias, and Trust

Future of Life Institute organized **The Asilomar Conference on Beneficial AI** conference which was held on January 5–8, 2017, at the Asilomar Conference Grounds in California. It was attended by more than 100 thought leaders and researchers in economics, law, ethics, and philosophy. The agenda of the conference was to address and formulate principles of beneficial AI. The outcome of the conference was the creation of a set of guidelines for AI termed *“the 23 Asilomar AI Principles”*.

The 23 Asilomar AI principles

Figure 4.2 describes the tenets of Asilomar principles:

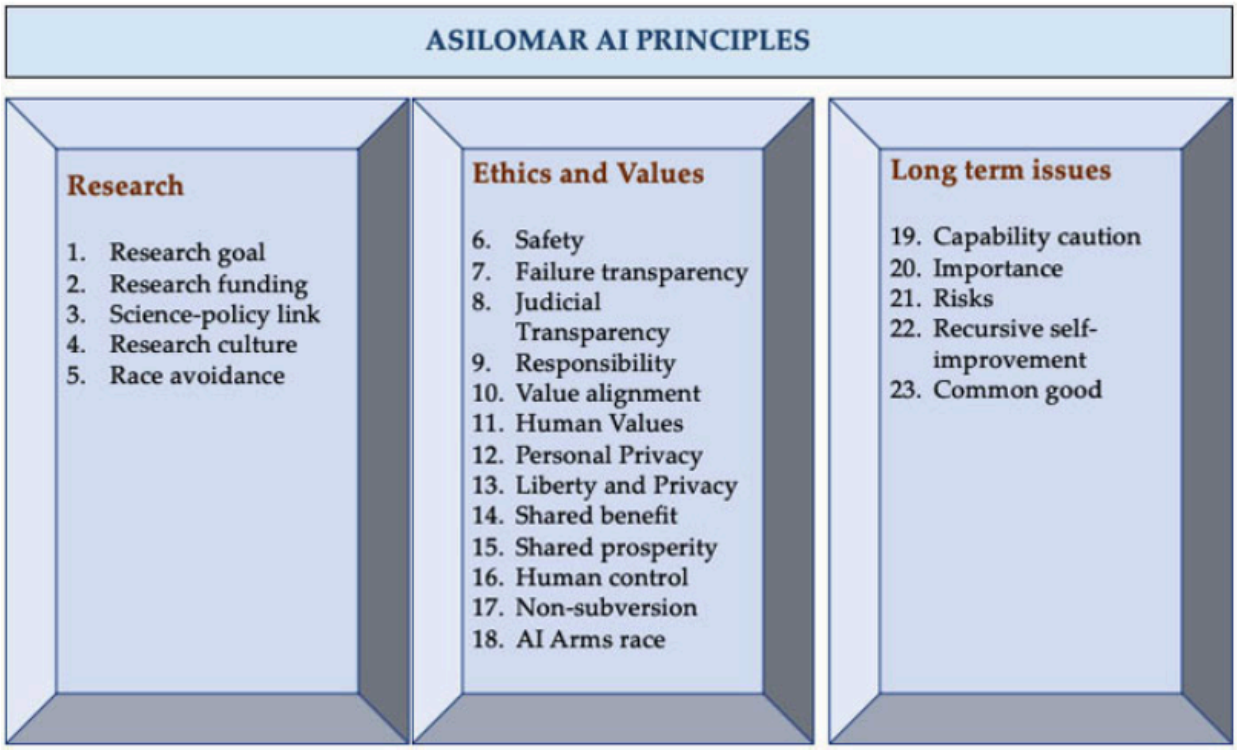

Figure 4.2: *Asilomar AI principles*

Research issues

Let us look at research issues further in the following:

Research goal:
The goal of AI research should be to create beneficial intelligence and not an undirected one.

Research Funding: Investments in AI should not only be accompanied by funding for research on ensuring its beneficial use, but also address questions related to economics, law, ethics, and social studies, such as:

- How to make future AI systems robust enough, so that they behave as per the aim of developing them without malfunctioning or getting hacked?
- How to achieve growth via automation while maintaining purpose?
- How could legal systems be updated to be more efficient, keeping pace with AI, and managing the risks associated with AI?

- What legal and ethical status should AI systems have?

Science-policy link:
AI researchers and policy-makers need to work hand in hand and have constructive and healthy exchanges of ideas and workings.

Research culture:
The working dynamics involving AI researchers and developers of AI must encourage a culture of cooperation, trust, and transparency.

Race avoidance:
There must be no compromise on safety standards due to teams developing AI systems not adhering to active cooperation.

Ethics and values

Now, we will discuss the ethics and values of Asilomar AI principles:

Safety:
AI systems should be designed such that the data, process, and systems in their entirety are safe, secure, and verifiable whenever applicable throughout their operations lifetime.

Failure transparency:
AI systems should be transparent enough such that it is possible to ascertain why any failure was caused, in case of one.

Judicial transparency:
In case of AI autonomous systems are involved in judicial decision-making, the system should provide a satisfactory explanation of the decision taken that is auditable by a competent human authority

Responsibility:
AI systems' designers and builders are to be held responsible in case of any moral implications of the use and misuse of these systems. Also, these stakeholders must take the opportunity to shape the implications.

Value alignment:
Goals and behaviors throughout the operational lifetime of highly autonomous AI systems should be assured to align with human values.

Human values:
AI systems should be designed and operated in such a way that they are aligned with human values such as ideals, dignity, rights, freedoms, and cultural diversity.

Personal privacy:
AI systems analyze the data generated by any human/user. This data must be transparently accessed, managed, and controlled by those who generate it.

Liberty and privacy:
Personal data fed to AI systems must not hinder the liberty, real or perceived, of humans.

Shared benefit:
AI technologies should benefit globally and empower communities, groups, or as many people as possible, irrespective of those who use it or those who do not.

Shared prosperity:
AI should benefit all humanity and help in economic prosperity.

Human control:
To accomplish human-chosen objectives, they control what to, whether to, and how much to delegate decision-making to AI systems.

Non-subversion:
Highly advanced AI systems must be designed and operated to improve rather than subvert the social and civic processes driving the health of society

AI Arms Race:
The race of owning Lethal autonomous weapons must be completely avoided in the interest of the human race.

Longer-term issues

The following are the long-term issues of Asilomar AI principles:

Capability caution:
No consensus has yet been established regarding future AI capabilities. Hence strong assumptions regarding the same should be avoided.

Importance:
Autonomous AI systems can have profound changes on the human race and life and hence should be planned for and managed with care and good intent.

Risks:
AI systems are expected to pose catastrophic or existential risks. As such, planning and mitigation efforts must match the degree of such impacts.

Recursive self-improvement:
Strict safety and control measures must be adopted for AI systems designed to recursively self-improve or self-replicate in a manner leading to rapidly increasing quality or quantity.

Common good:
AI systems that are defined as autonomous or super intelligent must be developed to achieve benefits for the entire humanity.

Employment and AI

Let us consider AI-powered applications and their impact on human employment.

Bias in recruitment

This is a classic example of how the data input itself can bring bias in the models. In a bid to *“out-recruit”* other technology firms, Amazon built an AI-based tool.

The models were trained to filter out top talents’ resumes. The model was trained based on data collected over 10 years. However, the data collected was tainted

as the majority of the candidates were men. This resulted in the AI model giving higher priority to male resumes while assigning low scores to the resumes that had participated in women's activities, such as "State- level Women's cricket team player". This was witnessed even when the names were anonymized.

Amazon gave up and disbanded the tool and the team after multiple failed attempts to make the program gender-neutral.

AI applications replacing humans

ChatGPT, an AI assistant, as also covered in previous sections is doing a faster and better job than humans, especially in content creation, making it very easy for enterprises to replace humans.

ChatGPT can also write entire computer code to program applications with accuracy. This is encouraging companies to use it instead of employing coders.

There are AI-powered applications that are capable of representing a lawyer in court for arguments. Thus, even the legal industry will see the impact of the adoption of AI technology and AI-powered systems.

The fear is thus not imaginary. Humans are losing their jobs with the advent of AI technology. It is anticipated that AI will have a huge impact on human employment but in certain fields only. AI, in its current state, is unable to replace humans, that is, to reach human intelligence.

Conclusion

AI-powered systems will have a deep impact on our society with the accelerated adoption and frequent use of AI applications for the smallest of tasks.

The impact of these AI-powered systems will not be limited to simple outputs, instead, these systems have the capacity to add on to the existing biases or the power to reduce the same. It, therefore, is very important for such systems to be based on ethics and adopt an unbiased approach.

One must spot issues within the data used for training and testing such systems. Also, make healthy arguments towards feature functionality and expected outcomes and apply policies of ethics and bias removals while designing such systems.

In the next chapter, we will be introduced to storytelling.