What is AI Ethics[¶](https://www.kaggle.com/code/raimondextervinluan/2023-kaggle-ai-report-ai-ethics-challenges#What-is-AI-Ethics)

Artificial Intelligence, or AI for short, has been conceptualized since the 1950s. The purpose of AI is to do things that a human can do but in a more efficient and accurate way. Another one is to make decisions based on certain parameters or environments.

Before we directly discuss on what *AI Ethics* is, first, let’s briefly discuss *Ethics* in general. Ethics is a branch of philosophy that always ask this question: “What is right? And what is wrong? Which choice should I make?” Ethics is our guide that defines conditions of a good choice or to figure out which of all the available options is the best one. But ethics does not stand on its own, it is mainly composed of three things: 1) Values, the things we hold to be good and therefore care about deeply. Some examples are family, equality, and justice. 2) Principles, which determine the acceptable ways of getting what we value. This is where we draw lines and make sure not to cross them. And lastly 3) Purpose, the reason for determining which principles and values to take on certain situations. But, even though there is a clear definition on what ethics is and what it is composed of, there is still no single set of ethical principles that everyone agrees on. To cope up with this, there are *some* general principles of ethics that are widely accepted such as “don’t do harm to others”, “be responsible in our actions”, and more.

So now we have briefly discussed on what *Ethics* is, how does this relate to AI Ethics? AI Ethics is a relatively new field that is concerned on how ethics can be combined with AI. Incorporating Ethics with AI means applying the generally accepted ethical principles and the three main things it is composed of. As AI develops, it is becoming more important to consider the ethical implications of how it is made and used.

Some Existing Ethical Frameworks or Guidelines

AI Ethics aims to ensure that AI is developed and used in a way that focuses to be beneficial to the people that it ought to serve and makes sure not to harm them in the process. Then, in order to achieve this, existing frameworks or guidelines have been implemented or is under review.

Kaggle’s Human-Centered Design for AI

From Kaggle’s “Intro to AI Ethics” course, there is something called Human-Centered Design for AI. Human-Centered Design, or HCD for short, is an approach to designing AI systems that serves people’s needs and involves people in every step of the design process. This will depend on the industry, resources, organization, and the people that the AI seeks to serve. There are six steps in HCD for AI.

The first step is “Understand people’s needs to define the problem” where the main point is to understand the pain points in people’s journeys. This can be done by observation, interviews, feedbacks, and focus groups. In this step, every team member should be involved to gain an understanding of the people they hope to serve.

Second step is “Ask if AI adds value to any potential solution”. In this step, there are some questions that needed to be address first before making the AI system. If there is a “no” to any of those questions, then maybe an AI solution may not be necessary or appropriate. These are the questions:

* Would people generally agree that what you are trying to achieve is a good outcome?
* Would non-AI systems - such as rule-based solutions, which are easier to create, audit, and maintain - be significantly less effective than an AI system?
* Is the task you are using AI for one that people would find boring, repetitive, or otherwise difficult to concentrate on?
* Have AI solutions proven to be better than other solutions for similar use cases in the past?

The third step is “Consider the potential harms that the AI system could cause”. Weigh the benefits of using AI against the potential harms throughout the design process. If the harms are likely to outweigh the benefits, then do not build the system. Uncover hidden privacy issues and determine whether privacy-preserving techniques may be appropriate. Thereby, always consider the impact of the AI system on users and on society.

Step four is “Prototype, starting with non-AI solutions”. Before developing an AI system, build first a non-AI prototype to see how people interact with it. Prototyping gives early information about what the users expect from your system and how to make their interactions more rewarding and meaningful.

Step five is “Provide ways for people to challenge the system”. The possible challenges that the people who use the system are the following:

* Requesting an explanation of how the AI arrived at its decision or answer
* Requesting a change in the information input
* Turning off certain features
* Reaching out to the product team on social media
* Taking some other action on the AI

The last and sixth step is “Build safety measures”. This is important because a safety measure protects users against harm. The kind of safety measure your system needs to depend on its purpose and on the types of harms it could cause. This can be done by reviewing the list of safety measure built into similar non-AI alternatives. Additionally, another safety measure is to let humans override and oversight the AI system.

These steps ensure that the AI system being built is aligned with what the people involved want and need.

Asilomar AI Principles

In 2017, the Asilomar Conference established guidelines called the Asilomar AI Principles. These principles promote responsible AI development, focusing on safety, security, and individual rights.

More than 5,000 individuals signed these principles, including 844 AI and robotics researchers. Among the prominent signatories include are Elon Musk and Stephen Hawking.

These principles recommend that AI should be developed where it reduces the risks of unintentional harm to humans. Additionally, AI should be developed where it benefits ALL members of society. Furthermore, the principles suggest to make policies that mitigate the negative effects of AI.

Another aspect of the principles is to promote transparency and accountability. Since AI systems are considered as black boxes, meaning that their decision-making process are difficult to understand, the principles call for AI systems to be designed in a way that is transparent and explainable.

**SOME** of the Key principles of the Asilomar AI Principles are:  
**Research Goal:** AI development should benefit all members of society and environment.  
**Safety:** AI should be designed to minimize or avoid the risk of causing unintentional harm to humans.  
**Transparency:** AI should be developed where its decisions can be easily understood and explained.  
**Privacy:** AI should be designed to protect the data it uses.  
**Fairness:** AI should be designed where it eliminates discrimination and bias.  
**Human Control:** Humans should be able to control AI at any given time or condition.  
**Responsibility:** The ones who develop AI should be held accountable for their impact in the society.

AI Ethics Lab’s TOOLBOX: Dynamics of AI Principles

This one is not a set of principles but rather a collection of all principles all over the world. This toolbox includes the following:

1. Map. Helps people to visualize where the principles are from and to see them in a summarized manner.
2. The Box. This is a tool to help people evaluate the ethical strengths and weaknesses of a particular AI technology.
3. Case studies. These are a set of case studies to see how the AI principles can be applied in practice.

To see this toolbox firsthand, visit this website: <https://aiethicslab.com/big-picture/>

Why is AI Ethics hard to incorporate?

While there may be numerous existing frameworks and guidelines, their implementation often encounters several hurdles. This may not be an extensive list of those hurdles but below, I will discuss some of it.

Bias is Hard to Get Rid Off

Bias will always be there. For us humans, we don’t have enough cognitive bandwidth to make every decision or assumption from scratch so we must make generalizations, or bias, as a starting point so that we can make the “best” decision as possible. If we cannot eliminate the bias on us, then how can we expect to eliminate the bias that is present on the AI if the data that the AI trains on came from us? Every time we train an AI, we unintentionally transfer our biases to them.

Values, Principles, and Purpose

I have defined earlier what these things are so now, I’ll discuss on why these things make the implementation of AI Ethics hard.

**Values such as justice, family, and equality.** Having values shape our ethical framework however, different people may prioritize or have a different set of values. This leads to disagreements and challenges when defining what is right or what decision to make in certain situations. This diversity of values makes it difficult to implement a AI Ethics.

**Principles serves as the means on how we achieve the values we hold.** In AI ethics, determining the acceptable ways of not crossing the line due to the fact that we also have different principles.

**Purpose helps us choose on which principles and values we should hold.** Determining a unified purpose for AI is challenging due to the same fact that we humans have different purposes.

Moral Uncertainty

Moral uncertainty refers to being indecisive on what is morally right or wrong in a certain situation. This happens when there are multiple moral choices to consider in making a decision.

So, why is this related to AI? Since there are multiple morals to consider, it is difficult to implement a universally accepted ethical guideline that can be applied to all situations that AI may encounter. For instance, lets say there is an ailment that only an AI System can cure. But the patient does not want to be treated that way. If the AI “decided” to respect the patient’s conscious decision to not be treated, then it will defy one of the key principles of the *Hippocratic Oath* which is to prioritize the well-being and health of patients above all else. On the other hand, if the AI “decided” to treat the patient, it will then violate the patient’s informed consent and privacy (for the reason that the patient’s data may be used for the treatment).

Moral Relativism

This one is a philosophical perspective that state ethical principles and judgements are not absolute or universally applicable. In addition, this states that ethics depend on cultural, societal, and individual views.

If ethics is relative, then how can AI know which ethical guidelines are appropriate to follow? Additionally, who will decide on which ethical guideline is applicable or to prioritize for certain situations?

**Regulatory Challenge**

Regulatory challenges do not only cover regulations; they also covers the problems with regulators and how regulations are implemented.

**Making AI regulations is not proactive.** Regulations are often delayed or lag behind innovations. One example of this is the fair lending laws in the USA are nearly 50 years old but are only starting to be modernized now in order to deal AI systems in practice. Moreover, the current pace of making regulations is out of sync with AI development. Top AI conference like NeurIPS reported 50% annual increase in their submissions, which the regulations of AI cannot keep up with. Additionally, technological regulations rely heavily on societal values. This results for regulations to be always in a state of reform as perspectives and issues change over time.

**AI is everywhere.** AI systems are used in many industries and fields such as healthcare, finance, and employment. The problem is that they often have their own regulations that are only applicable to them, and the AI systems made for specific industries only adhere to the AI regulations relevant to those industries. There is no single regulatory body to monitor and responsible for AI usage as a whole. This results to self-regulation, which carries an intrinsic risk of conflict of interest.

**Accountability for AI is hard.** AI often operate using complex algorithms and processes that are not easy to understand. Therefore, it is difficult to trace the decisions made by AI systems back to the specific actions or individuals. Additionally, since it is difficult to trace AI’s decision-making process, this may result to delays in action on what to do with the AI and it may continue causing harm for the people who are affected.

**Business Challenge**

In the business world, having a secret tool to be on edge with competitors is important. These secret tools make sure that the business or company will continue to grow and earn. With regards to using AI, companies treat them as the same. Keeping the inner workings of the AI systems a secret and share only limited information about the limitations and capabilities of their technology. This is for the reason to maintain the competitive edge and protect their intellectual property. The problem with this is that businesses or companies violate the existing guidelines or frameworks with regards to transparency and accountability.

**Limitations of the Asilomar AI Principles**

The Asilomar AI Principles may provide a useful guideline for the ethical use of AI, unfortunately, they still have some limitations.

**Not legally binding:** Although many AI developers signed these principles, they are not required by to comply.

**Focus:** The principles are more focused on the development and deployment, and not so much on the socioeconomic impact of AI on the society.

**Broad:** The principles are open to interpretations. Since the principles didn’t provide any specific, it makes the less effective.

**Assumption:** The principles assume that AI developers will act on good faith, which is unlikely. Some developers may intentionally develop AI systems for their own benefit only.