Turn Ethical Frameworks into Acceptable Actions

Ethical Principles and Frameworks

What is an ethical framework?

- Ethical framework as a set of guiding principles that govern how an individual or a company behaves and makes these decisions
- It is a set of ethical principles put together in such a way that is operationalized.

Major Schools of Philosophical Thought

- Utilitarianism
    - Most good for most people
    - Act utilitarianism
        - We judge every action by its effect on overall utility
    - Rule Utilitarianism
        - We follow a guideline of general rules that lead over time to optimal utility for the most amount of people
- Deontology (Duty Based)
    - Universal moral laws
- Virtue ethics
    - Right character comes before the right action

Montreal Declaration for Responsible Development of AI

- Well-Being
    - The development and use of artificial intelligence systems. Must permit the growth of the well-being of all sentient beings
- Respect for autonomy
    - The goal of increasing people's control over their lines and their surroundings
- Protection of privacy and intimacy
    - That privacy and intimacy must be protected from AI systems and by data acquisition and archiving systems
- Solidarity
    - The development of AI systems must be compatible with maintaining the bonds of human solidarity, among people and generations
- Democratic participation
    - AI systems must meet intelligibility, justifiability, and accessibility criteria. And must be subjected to democratic scrutiny, debate and control
- Equity
    - The people that need more gets more, while the people who need less get less
- Diversity and inclusion principle
    - The development of AI systems must be compatible with maintaining social and cultural diversity, and must not restrict the scope of lifestyle choices and personal experience
- Prudence principle

    - Every person in involved with the development of AI systems, they have to exercise caution by anticipating. As far as possible, the potential adverse consequences of AI system usage, and by taking the necessary precautions to avoid them
- Responsibility
    - The development and use of AI systems, they're not contributing to the diminishing of human responsibility
- Sustainable development
    - AI systems must be carried out, so as to ensure the strong environmental sustainability of the planet

Top 10 Principles for Ethical Artificial Intelligence

- Deman transparent AI systems
- Quit AI with Ethical black boxes
- AI serves the people and the planet
- Human in-command approach
- Ensure a genderless and unbiased AI
- Share the benefits of the AI systems
- Secure a just transition and ensure support for fundamental freedoms and rights
    - Ensure that people who are displaced by AI have a place to go
- Establish global governance mechanisms
- Don’t attribute responsibilities to robots
- Ban the AI arms race

Ethically Aligned Design

Align the creation of AIS systems with the value of society.

- Three pillars
    - Universal Human Values
        - AIS Systems should seek to promote and enhance human values and human rights
    - Political self-determination and data agency
        - AIS Systems should seek to enhance democracy and democratic values through the protection of user privacy
    - Technical dependability
        - Entrust these AIS systems to work on our behalf
- General Principles
    - Human Rights
        - AIS Systems should be created and operated to respect, and protect internationally recognized human rights
    - Wellbeing
        - Creators should adopt increased human well-being as primary success criterion for development
    - Data Agency
        - AI System creators should empower individuals with the ability to access and securely share their data
    - Effectiveness
    - Transparency

        - The basis of a particular AIS decision should always be discoverable and understandable
    - Accountability
        - AIS Systems should be created and operated to provide unambiguous and easy-to-understand rational decisions for human beings
    - Awareness of Misuse
    - Competence
        - AIS creators should specify and operators should actually understand how these AIS Systems work

The Asilomar AI Principles (abridged)

- Transparency
    - Failure transparency
        - Able to backtrace a failure
    - Judicial transparency
        - We can go back and change the AI decision if necessary
- Equal access to AI
    - AI technologies should benefit and empower as many people as possible
- Privacy
    - People should have the right to access, manage and control the data they generate, given AI systems' power to analyze and utilize that data.
    - Capability caution
    - That there being no consensus, we should avoid strong assumptions regarding the upper limits of feature AI capabilities

The Toronto Declaration (focus on Human Rights)

- International human rights law
    - Rights of quality and non-discrimination
- Duties of states
    - Obligations to ensure and protect human rights
        - Identify risks
        - Ensure transparency
        - Ensure accountability
        - Ensuring and enforcing oversight
- Responsibilities for private sector actors
    - Human rights due diligence
        - Identify potential discriminatory outcomes
        - Take effective action to prevent and mitigate discrimination and track responses
        - Be transparent about their attempts and efforts to identify, prevent, and mitigate discrimination
- Rights to an effective remedy
    - Reparations to victims of discrimination and outlining clean lines of accountability
    - Essentially the humans are responsible not the AI

Principles Shared by Major Ethical Frameworks

Protect Privacy

- Technical solutions
- More research
- Certifications and regulations

Accountability

- Causality
  - What caused harm
- Justice
  - Who to punish
- Reparations
  - Who pays

Transparency and Explainability

- Ability to see inside the AI system
- Describe in human terms
- Interpretability
  - See what's going on
  - Predict changes after modification
- Auditability
  - Ability to verify transparent/explainable

Fairness and Non-Discrimination

- Unbiased results
  - Independent of protected attributes
- Equitable access
  - To data and benefits of AI systems

Safety and Security

- Non-maleficence
  - Safe and secure operation
  - No foreseeable or unintentional harm

Human Control of Technology

- Meaningful human control
- Impermissible or unacceptable for an autonomous weapon or machine to keep on producing force, to operate without human supervision
- A person who simply clicks a button because a computer tells them to do so, without cognitive clarity or cognitive awareness, is usually insufficient to be considered meaningful human control.

Professional Responsibility

- Does AI augment professional knowledge or replace it

Promotion of Human Values