Ethics in Artificial Intelligence

By Jugal Kalita, PhD
Professor of Computer Science
Daniels Fund Ethics Initiative Ethics Fellow

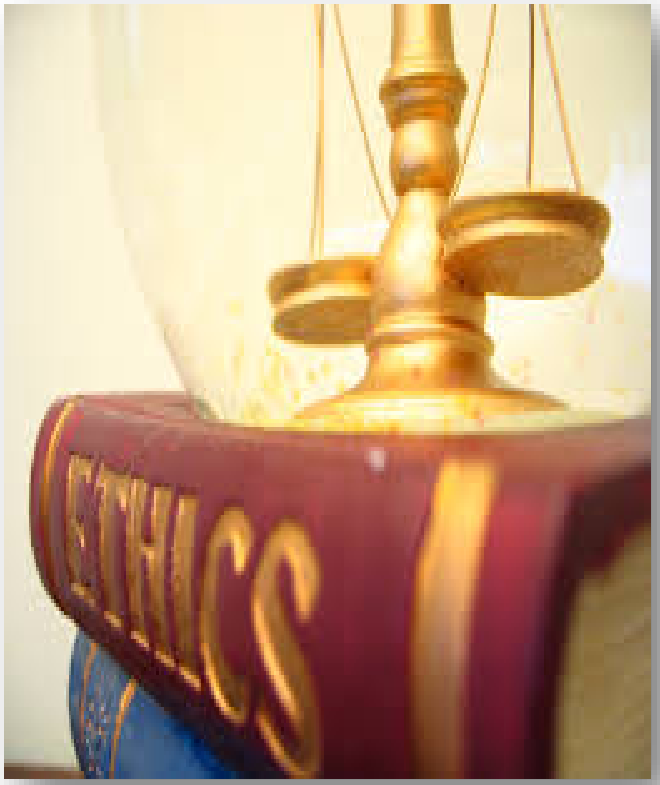
Sponsored by:

Ethics in Artificial Intelligence

- What are Ethics?
- What is Artificial Intelligence?
- What are Ethical issues in Artificial Intelligence?

Ethics

- Ethics or moral philosophy is a branch of philosophy that involves defining and discussing concepts of right and wrong conduct (Internet Encyclopedia of Philosophy)
- As a branch of philosophy, ethics investigates the questions "What is the best way for people to live?" and "What actions are right or wrong in particular circumstances?" (Source: Wikipedia)
- In practice, ethics seeks to resolve questions of human morality, by defining concepts such as good and evil, right and wrong, virtue and vice, justice and crime. Source: Wikipedia)



Types of Ethics

- *Meta-ethics*, dealing with theoretical meaning and reference of moral propositions: how we understand, know about, and what we mean when we talk about what is right and what is wrong.
- *Normative ethics*, dealing with practical means of determining a moral course of action: Virtue ethics, Deontology, Utilitarianism, etc.
- *Applied ethics*, dealing with what a person is obligated (or permitted) to do in a specific situation or a particular domain of action: Bioethics, Business ethics, **Machine ethics**, Military ethics, etc.

Machine Ethics

- Machine ethics (or machine morality, computational morality, or computational ethics) is a part of the ethics of artificial intelligence concerned with the moral behavior of artificially intelligent beings (Moor, 2006. "The Nature, Importance and Difficulty of Machine Ethics, IEEE Intelligent Systems).

- Mitchell Waldrop in the 1987 AI Magazine article "A Question of Responsibility":

  "However, one thing that is apparent … is that intelligent machines will embody values, assumptions, and purposes, whether their programmers consciously intend them to or not. Thus, as computers and robots become more and more intelligent, it becomes imperative that we think carefully and explicitly about what those built-in values are. Perhaps what we need is, in fact, a theory and practice of machine ethics, in the spirit of Asimov's three laws of robotics. "

  (AAAI Presidential Panel on Long-Term AI Futures 2008-2009 Study)

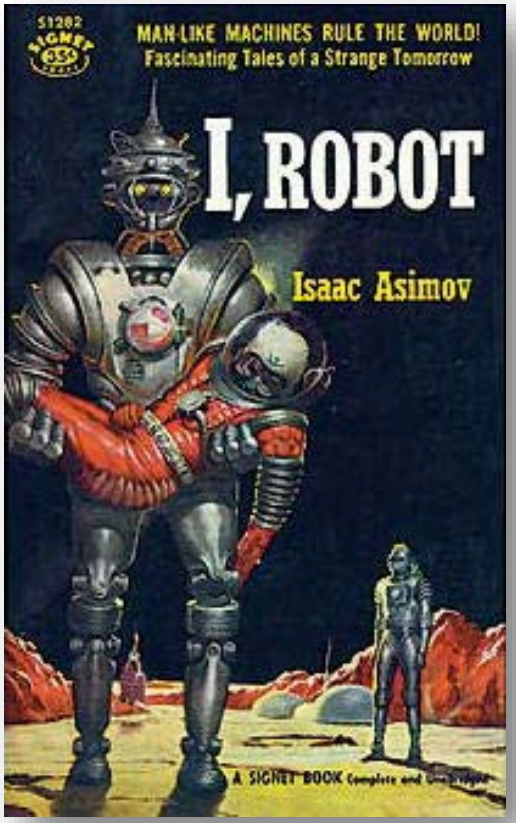
Artificial Intelligence

- Artificial Intelligence is a field of study concerned with the design and construction of intelligent agents (Poole et al. 1998).
- An agent may be a software agent, or a hardware agent controlled by software.
- AI, as a field of study, is about 60 years old.

“Laws” of Robotics

The Laws of Robotics are a set of rules devised by the science fiction author Isaac Asimov. The rules were introduced in his 1942 short story "Runaround” in the book “I, ROBOT”.

- **Rule 0**: A robot may not harm humanity, or, by inaction, allow humanity to come to harm.
- **Rule 1**: A robot may not injure a human being or, through inaction, allow a human being to come to harm.
- **Rule 2**: A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
- **Rule 3**: A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.



Types of AI Agents

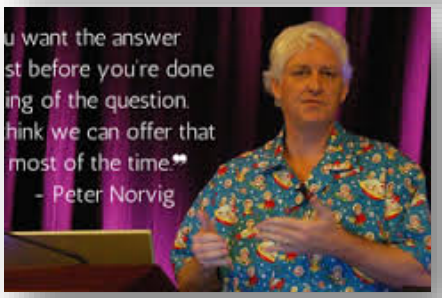
- Strong AI or General Artificial Intelligence: A machine with consciousness, sentience and mind or a machine with the ability to apply intelligence to any problem, rather than just one specific problem.
- Weak AI or narrow AI: A non-sentient artificial intelligence that is focused on one narrow task.
- Weak or "narrow" AI is a present-day reality. Software controls many facets of daily life and, in some cases, this control presents real issues. One example is the May 2010 "flash crash" that caused a temporary but enormous dip in the market (Ryan Calo, Center for Internet and Society, Stanford Law School, 2011).

Strong AI

- The chief problem of building human-like artificial intelligence: the human mind is inexorably linked to the body. (Dreyfus, On the Internet, 2009).
- Isolating the mind from the body is a fundamentally flawed approach to this kind of AI development because it ignores the aspect of the mind that develops solely because humans have bodies.
- Humans have a perception of what it is to feel relaxed, or to feel pain, or to feel temperature, because of interaction between the body, the mind, and the environment.
- A piece of software, since it is physically formless and has no way of interacting with the environment, can only form abstractions about these feelings, most likely based on interpretations of numerical values (i.e. 90 degrees is hot, 30 degrees is cold). (Dreyfus)
- Strong AI: can reason, make judgments, plan activities, learn, communicate, conscious, self-aware, sentient, sapient (https://www.youtube.com/watch?v=5nwUJnlvjCc).

Stephen Hawking + Elon Musk Endorse Principles of AI (1 Feb 2017)

- Safety: AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible.
- Failure Transparency: If an AI system causes harm, it should be possible to ascertain why.
- Value Alignment: Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.
- Human Values: AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity.
- Personal 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.
- Liberty and Privacy: The application of AI to personal data must not unreasonably curtail people's real or perceived liberty.
- Shared Benefit: AI technologies should benefit and empower as many people as possible.
- **Shared Prosperity:** The economic prosperity created by AI should be shared broadly, to benefit all of humanity.
- Human Control: Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives.
- AI Arms Race: An arms race in lethal autonomous weapons should be avoided.



Some Ethical Issues in AI

- Unemployment: AI may lead to massive amounts of job loss, e.g., in the trucking industry, in the taxicab industry.
- Inequality: How should the wealth generated by AI-based industries be distributed? Owners of these companies are likely to have huge paydays exacerbating income inequality.
- AI Bias or AI Racism: Although the AI programs can learn and adapt, they still are likely to show the biases of the developers. E.g., a program that scores people for mortgage may score African Americans low based on a chance pattern it finds.
- Security of AIs: Many AI programs are likely to be able to do good things as well bad and nefarious. The AI programs or devices must be secured properly so that they don't fall into the wrong hands.
- (source: Word Economic Forum and other source)

Some Ethical Issues in AI

- Questions of Life and Death: AI-powered vehicles such as cars and trucks may need to make decisions regarding situations that may lead to questions of life and death. AI-powered medical systems and military systems have to make such decisions regularly.
- Weaponization of AI: AI based devices, vehicles, drones, etc., have the potential to act as weapons with some amount of intelligence. E.g., a small drone may fly into a house, go through the rooms and find an “enemy” to shoot and kill.

Some Ethical Issues in AI

- Humans in Control: AI-powered entities need to remain under the control of responsible humans. We have done so with bigger and stronger animals, and objects like airplanes, using physical controls like cages, keys and weapons; or using cognitive tools like training.
- Rights of AIs: AIs are still simple and mostly disembodied programs. But, as AIs become more complex and start to have physical possibly humanio shapes, and more numerous, what rights should they have?

University of Colorado
Colorado Springs

University of Colorado

Boulder | Colorado Springs | Denver | Anschutz Medical Campus