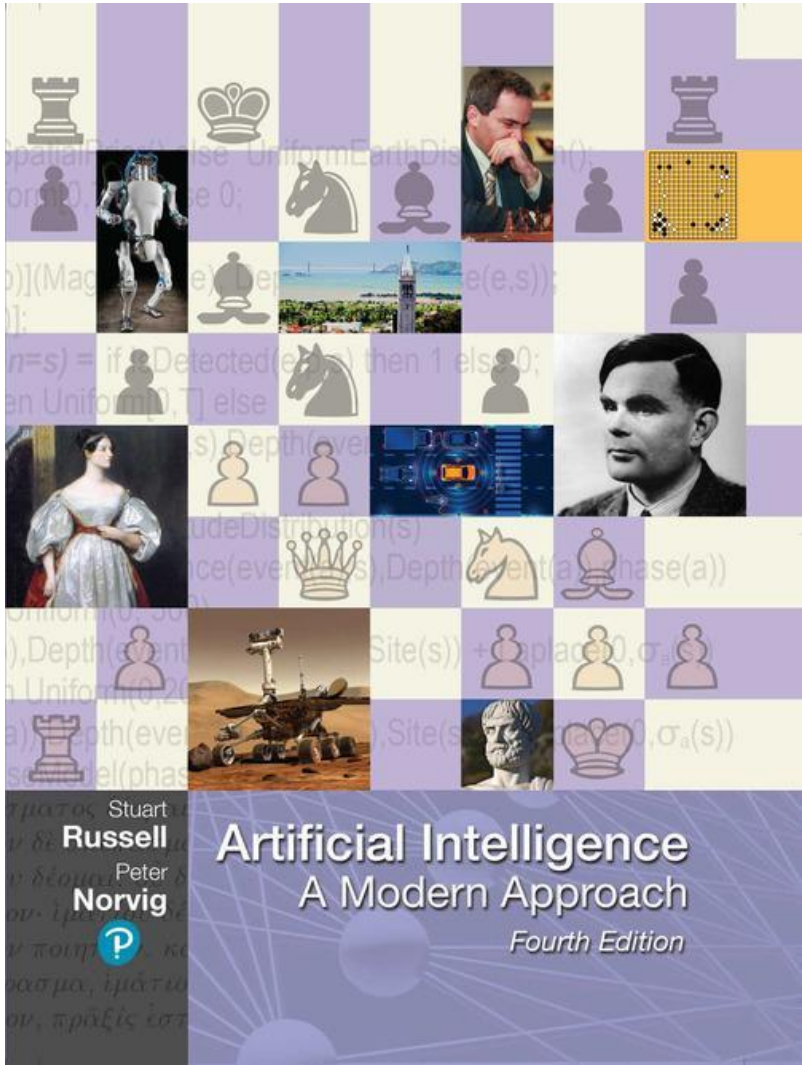
Artificial Intelligence: A Modern Approach

Fourth Edition



Chapter 27

Philosophy, Ethics, And Safety Of AI

Outline

- The Limits of AI
- Can Machines Really Think?
- The Ethics of AI

The Limits of AI

Philosopher John Searle (1980):

- **weak AI**: the idea that machines could act as if they were intelligent
- **strong AI**: the assertion that machines that do so are actually consciously thinking (not just simulating thinking)

The argument from informality

Turing's "argument from informality of behavior" says that human behavior is far too complex to be captured by any formal set of rules

Good Old-Fashioned AI (GOFAI)

- simplest logical agent design
- qualification problem: difficult to capture every contingency of appropriate behavior in a set of necessary and sufficient logical rules
- Dreyfus's strongest arguments is for situated agents rather than disembodied logical inference engines
- **embodied cognition** approach claims that it makes no sense to consider the brain separately
  - cognition takes place within a body, which is embedded in an environment [**We are good at frisbee, bad at logic**.]

The Limits of AI

The argument from disability

- The "argument from disability" makes the claim that "a machine can never do *X*."

- Turing's lists of *X*:

Be kind, resourceful, beautiful, friendly, have initiative, have a sense of humor, **tell right from wrong,** make mistakes, fall in love, enjoy strawberries and cream, make someone fall in love with it, learn from experience, use words properly, be the subject of its own thought, have as much diversity of behavior as man, do something really new.

- Some of these are rather easy to be replicated by AI. However some are not possible

- Overall, programs exceed human performance in some tasks and lag behind on others.

- The one thing that it is clear they can't do is be exactly human.

The Limits of AI

The mathematical objection

Turing (1936) and Gödel (1931) proved that certain mathematical questions are in principle unanswerable by particular formal systems.

Gödel sentence $G(F)$ with the following properties:

- $G(F)$ is a sentence of F, but cannot be proved within F.
- If F is consistent, then $G(F)$ is true.

Philosophers such as J. R. Lucas (1961) have claimed that this theorem shows that machines are **mentally inferior to humans**,

- machines are **formal systems** that are limited by the incompleteness theorem
- cannot establish the truth of their own Gödel sentence
- Problems with Lucas' claim:
  - Example sentence which **cannot consistently assert** by human else contradiction:

    Lucas cannot consistently assert that this sentence is true.
  - **No entity**—human or machine—can prove things that are impossible to prove
  - **incompleteness theorem** technically applies only to **formal systems** that are powerful enough to do arithmetic.

The Limits of AI

Measuring AI

- whether machines can pass a behavioral test, which has come to be called the **Turing test**
- The test requires a program to have a conversation (via typed messages) with an interrogator for five minutes

- ELIZA program and Internet chatbots such as MGONZ and NATACHATA

- Eugene Goostman fooled 33% of the untrained amateur judges in a Turing test

- AI researchers who crave competition are more likely to concentrate on playing chess or Go or StarCraft II, or taking an 8th grade science exam, or identifying objects in images. [**See NLPProgress]**

Can Machines Really Think?

Some philosophers claim that a machine that acts intelligently would not be actually thinking, but would be only a simulation of thinking

Turing argues the **polite convention** that everyone and machine think.

John Searle rejects the polite convention

The Chinese room

- A human, who understands only English, inside a room that contains a rule book, written in English, and various stacks of paper
- Pieces of paper containing indecipherable symbols are slipped under the door to the room
- The human follows the instructions in the rule book, finding symbols in the stacks, writing symbols on new pieces of paper, rearranging the stacks, and so on
- passed back to the outside world
- it is given that the human does not understand Chinese
- computer are in essence doing the same thing, so therefore computers generate no understanding

The Ethics of AI

Given that AI is a powerful technology, we have a moral obligation to use it well, to promote the positive aspects and avoid or mitigate the negative ones.

Positive aspects examples

- AI can save lives through improved medical diagnosis, new medical discoveries, better prediction of extreme weather events [**Self driving cars 30,000 fatalities / year - the perfect is the enemy of the good.**]

- AI can improve lives, Microsoft's AI for Humanitarian Action program applies AI to recovering from natural disaster

- AI applications in crop management and food production help feed the world

The Ethics of AI

Negative aspects example

- **Lethal autonomous weapons**

The UN defines a lethal autonomous weapon as one that locates, selects, and engages (i.e., kills) human targets without human supervision.

Israel's Harop missile is a "loitering munition" with a ten-foot wingspan and a fifty-pound warhead. It searches for up to six hours in a given geographical region for any target that meets a given criterion and then destroys it.

Autonomous weapons have been called the "third revolution in warfare" after gunpowder and nuclear weapons. Their military potential is obvious

The debate over autonomous weapons includes **legal, ethical and practical aspects.**

Legal: requires the possibility of discriminating between combatants and non-combatants, the judgment of military necessity for an attack, and the assessment of proportionality between the military value of a target and the possibility of collateral damage.

The Ethics of AI

Negative aspects example

- **Lethal autonomous weapons**

Ethical: some find it simply morally unacceptable to delegate the decision to kill humans to a machine.

More than 140 NGOs in over 60 countries are part of the Campaign to Stop Killer Robots, and an open letter organized in 2015 by the Future of Life Institute organized an open letter was signed by over 4,000 AI researchers and 22,000 others

Reliability: a very serious concern for military commanders, who know well the complexity of battlefield situations. Cyberattacks against autonomous weapons could result in friendly-fire casualties

Practical: the scale of an attack that can be launched is proportional to the amount of hardware one can afford to deploy.

AI is a **dual use technology**: AI technologies that have peaceful applications can easily be applied to military purposes

The Ethics of AI

Surveillance, security, and privacy

- As of 2018, there were as many as 350 million surveillance cameras in China and 70 million in the United States.
- As more of our institutions operate online, more vulnerable to cybercrime and cyberterrorism. Attackers can use automation to probe for insecurities and they can apply reinforcement learning for phishing attempts and automated blackmail
- Defenders can use unsupervised learning to detect anomalous incoming traffic patterns and various machine learning techniques to detect fraud
- More data on us is being collected by governments and corporation
- In the U.S., the Health Insurance Portability and Accountability Act (HIPAA) and the Family Educational Rights and Privacy Act (FERPA) protect the privacy of medical and student record
- **De-identification**: eliminating personally identifying information (such as name and social security number) so that medical researchers can use the data to advance the common good
  - Federated learning
- **Secure aggregation:** central server doesn't need to know the exact parameter value from each distributed user

The Ethics of AI

Fairness and bias

- machine learning models can perpetuate societal bias

- Designers of machine learning systems have a moral responsibility to ensure that their systems are fair

- six of the most commonly-used concepts for fairness: [**See Rae cite**]
  - Individual fairness
  - Group fairness
  - Fairness through unawareness
  - Equal outcome
  - Equal opportunity
  - Equal Impact

- **COMPAS** is a commercial system for recidivism (re-offense) scoring. It assigns to a defendant in a criminal case a risk score, which is then used by a judge to help make decisions
  - does not achieve equal opportunity: the proportion of those who did not re-offend but were falsely rated as high-risk was 45% for blacks and 23% for whites

The Ethics of AI

Fairness and bias

- **sample size disparity** can lead to biased results.
- In most data sets there will be fewer training examples of minority class
- Machine learning algorithms give better accuracy with more training data, so that means that members of minority classes will experience lower accuracy
- A constrained model may not be able to simultaneously fit both the majority and minority class
- Bias can also come into play in the software development process
- De-bias the data: over-sample from minority classes to defend against sample size disparity

The Ethics of AI

Set of best practices

- Make sure that the software engineers talk with social scientists and domain experts to understand the issues and perspectives, and consider fairness from the start.
- Create an environment that fosters the development of a diverse pool of software engineers that are representative of society.
- Define what groups your system will support: different language speakers, different age groups, different abilities with sight and hearing, etc.
- Optimize for an objective function that incorporates fairness.
- Examine your data for prejudice and for correlations between protected attributes and other attributes.
- Understand how any human annotation of data is done, design goals for annotation accuracy, and verify that the goals are met.
- Don’t just track overall metrics for your system; make sure you track metrics for subgroups that might be victims of bias.
- Include system tests that reflect the experience of minority group users.
- Have a feedback loop so that when fairness problems come up, they are dealt with

The Ethics of AI

Trust and transparency

- People need to be able to trust the systems they use

- Engineered systems must go through a verification and validation (V&V) process
  - Verification means that the product satisfies the specifications
  - Validation means ensuring that the specifications actually meet the needs of the user and other affected parties

- Certification and safe standards, ISO in other industries

- The AI industry is not yet at this level of clarity, although there are some frameworks in progress, such as IEEE P7001, a standard defining ethical design for artificial intelligence and autonomous systems

- **Transparency**: consumers want to know what is going on inside a system, and that the system is not working against them, whether due to intentional malice, an unintentional bug, or pervasive societal bias that is recapitulated by the system

The Ethics of AI

Trust and transparency

An AI system that can explain itself is called **explainable AI (XAI).**

A good explanation has several properties:

- it should be understandable and convincing to the user
- it should accurately reflect the reasoning of the system
- it should be complete,
- it should be specific in that different users with different conditions or different outcomes should get different explanations.

The future of work [Aristotle *Politics*]

- an immediate reduction in employment when an employer finds a mechanical method to perform work previously done by a person

- More automation with physical robots, first in controlled warehouse environments, then in more uncertain environments, building to a significant portion of the marketplace by around 2030. [2**% farmers in 2010, 25% played FarmVille**]

- the ratio between workers and retirees changes. In 2015 there were less than 30 retirees per 100 workers; by 2050 there may be over 60 per 100 workers

- problems due to the pace of change [**ATM example: eliminate tasks, not jobs**]

The Ethics of AI

Robot rights

- if robots can feel pain, if they can dread death, if they are considered "persons," then the argument can be made that they have rights and deserve to have their rights recognize
- If robots have rights, then they should not be enslaved, and there is a question of whether reprogramming them would be a kind of enslavement
- Another ethical issue involves voting rights: a rich person could buy thousands of robots and program them to cast thousands of votes—should those votes count?
- **Ernie Davis** argues for avoiding the dilemmas of robot consciousness by never building robots that could possibly be considered conscious

The Ethics of AI

AI Safety

- design a robot to have low impact, instead of just maximizing utility, maximize the utility minus a weighted summary of all changes to the state of the world.
- Victoria Krakovna (2018) has cataloged examples of AI agents that have gamed the system, figuring out how to maximize utility without actually solving the problem that their designers intended them to solve.
- Genetic algorithm operating in a simulated world was supposed to evolve fast-moving creatures but in fact produced creatures that were enormously tall and moved fast by falling over.
- Designers of agents should be aware of these kinds of specification failures and take steps to avoid them.
- need to be very careful in specifying what we want, because with utility maximizers we get what we actually asked for. The value alignment problem

Summary

- Philosophers use the term weak AI for the hypothesis that machines could possibly behave intelligently, and strong AI for the hypothesis that such machines would count as having actual minds (as opposed to simulated minds

- AI is a powerful technology, and as such it poses potential dangers, through lethal autonomous weapons, security and privacy breaches, unintended side effects, unintentional errors, and malignant misuse. Those who work with AI technology have an ethical imperative to responsibly reduce those dangers.

- AI systems must be able to demonstrate they are fair, trustworthy, and transparent

- There are multiple aspects of fairness, and it is impossible to maximize all of them at once. So, a first step is to decide what counts as fair

- Automation is already changing the way people work. As a society, we will have to deal with these changes.