Ethics of AI

A guide to AI ethics

Since Alan Turing invented computation in the 1950's, humanity has held high hopes for the power of computers and artificial intelligence (AI). AI is anticipated to provide significant and varied benefits to society, from increased efficiency and productivity to addressing several challenging global issues such as climate change, poverty, disease, and conflict.

AI technologies shape our societies and significantly influence our daily lives. Simultaneously, various legal and societal issues have exposed the potential of these technologies to create negative effects. Algorithms can amplify existing biases, discriminate, compromise our security, manipulate us, and lead to lethal consequences.

For these reasons, people must explore the ethical, social, and legal aspects of AI systems. There is a widespread call for an ethical approach to AI. How can we develop and use this technology in an ethically acceptable and sustainable manner? What ethical and moral principles should we adopt and uphold?

In this lesson, we will explore the ethical issues surrounding contemporary AI, discuss their philosophical background, and interpret them in relation to computer science and other fields. The goal is to develop skills in AI ethical thinking.

What is AI?

Two Areas

Artificial intelligence is a broad term encompassing various methods that enable computers to exhibit intelligent behavior.[1] While there is no universally accepted definition of AI, the capacity to perform tasks independently and to learn to enhance performance are key traits of AI.

Machine learning is a crucial area of AI. It involves algorithms that autonomously learn to make decisions or sort data. Supervised and unsupervised learning depend on data, while reinforcement learning involves algorithms learning to create sequences.

What is AI ethics?

AI ethics is a subfield of applied ethics. Today, AI ethics is viewed as a branch of technology ethics focused on robots and other artificial intelligence systems. It addresses how developers, manufacturers, authorities, and operators should act to minimize the ethical risks that may emerge from AI in society, whether arising from design, improper use, or deliberate misuse of the technology.

These concerns can be divided into three time frames as follows:

- **Immediate**: here-and-now questions regarding security, privacy, or transparency in AI systems
- **Medium-term:** concerns about, for instance, the impact of AI on the military use, medical care, justice, and educational systems
- **Longer-term:** concerns about the fundamental ethical goals of developing and implementing AI in society

[1] https://tokezi.com/how-artificial-intelligence-can-contribute-to-combating-climate-change/

From Machine Ethics To The Ethics Of AI

For a long time, AI ethics primarily involved machines and robotics, focusing on the ethical codes of artificial moral agents. This field of research explores scenarios where machines could eventually be responsible for ethically significant decisions, potentially qualifying them as ethical or autonomous moral agents. In contrast, animals are typically not regarded as moral agents. We tend not to judge a squirrel's behavior as correct or wrong, nor do we assume they can recognize the difference.

Machine learning and robotics encompass everything from the development of ethically responsive autonomous vehicles to the design of ethical codes for moral autonomous agents. Isaac Asimov (1942) famously proposed "three laws of robotics" that would guide the moral action of machines:

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

Nowadays, AI ethics is a more general field closer to engineering ethics: We don't have to assume the machine is an ethical agent to analyze its ethics. Research in AI ethics ranges from reflections on how ethical or moral principles can be implemented in autonomous machines to empirical analyses of how trolley problems are solved, systematic analyses of ethical principles such as fairness, and critical evaluations of ethical frameworks.

A Framework For AI Ethics

Traditionally, technological development has revolved around functionality, usability, efficiency, and reliability. However, AI technology needs a broader discussion of its societal acceptability. It impacts moral (and political) considerations. It shapes individuals, societies, and their environments in ways that have ethical implications.

The interpretation of ethically relevant concepts can evolve with technology (consider what "privacy" meant before social media). Moreover, when new technologies emerge, users often utilize them for purposes other than those originally intended.[3]. This reshapes the ethical landscape and compels us to reflect on and analyze the ethical foundations of technology continuously.

Ethical frameworks

Ethical frameworks aim to build consensus around values and norms that a community can adopt—whether that community comprises individuals, citizens, governments, businesses in the data sector, or other stakeholders.

[2] Check out the I, Robot movie

[3] Examples: Napster, AI image creators and pornography

Various organizations have participated in the development of an ethical framework for AI. Naturally, their views differ in some respects, but there has also been an emerging consensus among them. According to a recent study, AI ethics has rapidly converged on five principles.[4.]

- non-maleficence or beneficence
- responsibility or accountability
- transparency and explainability
- justice and fairness
- respect for various human rights, such as privacy and security

The five principles of AI ethics answer different questions and focus on different values:

1. Should we use AI for good and not to cause harm? (the principle of beneficence/ non-maleficence)
2. Who should be blamed when AI causes harm? (the principle of accountability)
3. Should we comprehend what AI does and why? (the principle of transparency)
4. Should AI be fair or non-discriminative? (the principle of fairness)
5. Should AI respect and promote human rights? (the principle of respecting basic human rights)

Lastly, I want to emphasize that when discussing AI and its social implications, AI ethics is paramount. However, there are additional theoretical frameworks for examining ethical codes for algorithmic and data-driven systems. For example, questions about the social implications of AI arise in fields such as algorithmic culture, bias, and media studies, among many others. Similarly, the cognitive and psychological aspects of human-machine interaction influence the discourse on the suitable ethical framework for AI. In simple terms, AI ethics encompasses much more than just data or algorithm ethics.

What Should We Do?

What do the principles of beneficence (do good) and non-maleficence (do no harm) mean for AI, and how do they relate to the concept of the "common good?[5]

The Principle Of Beneficence

The principle of **beneficence state**s, "Do good," while the principle of **non-maleficence assert**s, "Do no harm." Though these two principles may seem similar, they represent distinct concepts. Beneficence promotes the development of beneficial AI ("AI should be created for the common good and the benefit of humanity"). At the same time, non-maleficence focuses on the negative consequences and risks associated with AI.

AI ethics has generally been concerned primarily with the principle of non-maleficence. Discussions have focused mostly on how developers, manufacturers, authorities, or other stakeholders should minimize the ethical risks—discrimination, privacy protection, and physical and social harms—that can arise from AI applications. Often, these discussions are expressed in terms of intentional misuse, malicious hacking, technical measures, or risk-management strategies.

[4] Jobin et al. 2019

[5] We discussed these terms in Lesson 10 as part of the idea of bioethics

Critics argue that the focus on non-maleficence turns ethics into a quest for technical solutions to technical problems. Moral dilemmas are viewed as challenges that can be addressed solely through technical "fixes" or effective design. The broader ethical and societal context surrounding these technical systems is often overlooked. As a result, many important issues concerning the control, governance, and societal aspects of AI are neglected.

As a result, deep and difficult ethical problems are oversimplified and unanswered. One of the questions is the problem of the "common good." What, exactly, does that mean?

What does accountability actually mean, and how does it apply to AI ethics? What do moral agency and responsibility mean, and what is the difficulty of assigning blame?

What is Accountability?

Accountability means being responsible or answerable for a system, its behavior, and its potential impacts. It is an acknowledgment of responsibility for actions, decisions, and products.

Responsibility can be legal or moral (ethical). **Legally**, an actor is responsible for an event when a legal system is liable to penalize that actor for that event. **Morally**, an actor is responsible for an act if they can be blamed for the action. Moral and legal responsibility are different things. They do not always coincide; an agent can be legally responsible even if they are not morally responsible, and vice versa. In this course, we will focus only on the moral aspects of responsibility.

In AI ethics, there are three different senses or dimensions of accountability. They point to different means of action, including:

- The question of determining the responsibility – which individuals (or groups) are accountable for the impact of algorithms or AI? Who is responsible for what effect within the overall socio-technical system?
- A feature of the societal system that develops, produces, and uses AI
- A feature of the AI system itself

Who should be blamed – and for what?

In ethics, accountability is closely linked to "moral agency." A moral agent is defined as "an agent who is capable of acting with consideration of right and wrong." Importantly, only moral agents can be held morally responsible for their actions.

Actions And Omissions

Philosophically, a moral agent is primarily accountable for their actions, referred to as "acts." At times, agents are also responsible for their inaction, known as "omissions."[6] Therefore, if I kill someone, I am responsible for that act. If I merely let someone die, I am still accountable for my inaction (omission), even if I did not actively kill anyone. Actions and omissions hold different moral weights; it is less wrong to omit something than to commit an act. Killing someone is worse than allowing them to die. However, this does not make omissions morally acceptable. We cannot be held accountable for everything we fail to do; instead, we are responsible only for the actions we deliberately and knowingly choose to perform or neglect.

[6] Lesson 2, Morally Obligatory and Morally Permissible

Autonomy

Philosophically, moral responsibility necessitates 1) moral autonomy and 2) the ability to assess the consequences of actions. “Moral autonomy” refers to the agent’s capacity to apply the moral code to oneself in a self-governing manner. Additionally, autonomy requires:

- The capacity to rule oneself without manipulation by others and the ability to act without external or internal constraints
- The authenticity of the desires (values, emotions, etc.) that move someone to act
- Sufficient cognitive skills – meaning an agent must be able to evaluate, predict, and compare consequences of their actions, and also to estimate motives that drive action by using ethically meaningful criteria

The Problem Of Individuating Responsibilities

Accountability is often viewed as a legal and ethical duty for individuals or organizations to take responsibility for their use of AI systems and to disclose the outcomes transparently. This concept assumes a “power relationship," determining who holds control and who is to be blamed.

However, it has been notoriously difficult to establish specific criteria for identifying, directing, and defining responsibilities. Ongoing debates on these questions are occurring in many countries.

Why is it so difficult to set criteria on who is responsible?

- **Firstly**, the quality of responsibilities differs. An actor is responsible for a specific action or omission, but the responsibility quality depends on the stakeholder. Thus, although choosing an action may entail responsibility, the quality of that responsibility also depends on your properties. Intelligent technologies complicate this further.
  As we delegate more decision-making tasks and functions to algorithms, we also shape the structures of decision-making.[7]. AI is augmenting our intelligence by giving us more computational power, allowing better predictions, and enhancing our sensory apparatus. Humans and machines become cognitive hybrids. They cooperate cognitively (thinking) and epistemically (knowledge), both at the individual and collective levels, creating systemic properties.
  It is often thought that staying “in-the-loop” or “on-the-loop” is sufficient—meaning that at some point during decision-making, a human individual would be able to monitor or intervene in the artificial system. However, as algorithms enter decision-making processes, such as in public-sector governance, collective decision-making can become very complex and highly distributed. It may be difficult to identify and address the factors in a way that guarantees that a human stays in/on-the-loop.
- **Secondly**, technology can also take on a persuasive role by influencing and controlling people. A classic example is the beeping sound of seat belts. In many cars, if the seat belts are unfastened, it triggers a continuous beeping sound. This can be seen as a form of control—essentially coercion. The driver can only silence the sound by fastening the belt. Modern algorithmic applications are increasingly incorporating similar features; they propose, suggest, and restrict options.

[7] War Games Movie

An action is considered voluntary only if it is done intentionally, meaning the person acting is “in control’ and free from external influences. Is the driver free from such influences if the seat belt system compels him to react to the beeping sound? Are we truly free from control when algorithms determine which pictures we see on dating sites or what music we listen to? What precisely is the difference between algorithmic suggestion, control, and manipulation?

Persuasive technology must inherently adhere to the principle of voluntariness to ensure autonomy. Algorithms complicate this matter because voluntariness implies a sufficient comprehension of the specific technology's use. But what does it mean to "understand," and what constitutes a sufficient degree of understanding, really? What is the correct interpretation of “understandability” – "transparency," "explainability,” or “auditability”? How much and what should a user realistically grasp about the technology? When can one accurately determine whether to engage with that particular technology?

Should We Know How AI Works

Why is transparency in AI important? What significant issues does it affect, and what are some risks associated with transparency in AI systems?

Transparency in AI

The principle of transparency

Consider a facial recognition system that is used for airport security. Usually, it operates flawlessly, but one day, it begins misidentifying individuals as potentially dangerous. As a result, several innocent people are detained. Is it essential to understand why the system made these errors? Should we be able to clarify its mistakes? Why does this matter?

Some contemporary machine learning systems are called “black box” systems, which means we cannot see how they operate. This “opacity,” or lack of visibility, can pose a problem when we use these systems to make decisions that affect individuals.

Individuals have a right to know how critical decisions are made, such as who is accepted for a loan application, paroled, and hired. This has led many to call for “more transparent AI.”

Transparency in AI

Transparency is a system characteristic that permits access to specific information about its inner workings. However, the nature of the information provided, as well as its ethical significance, largely depends on the ethical questions we seek to address. Transparency is ethically neutral and is not an ethical concept; instead, it represents an ideal. It can take various forms and can aid in tackling underlying ethical questions. In this context, transparency is pertinent to at least three key issues:

1. **The justification of decisions.** Good governance in public and private sectors requires that decisions be made without arbitrariness. This principle applies to any decision-making process that has ethical or legal implications for individuals. Non-arbitrariness entails having access to justifications that explain “why this decision was made, and on what basis?” Additionally, particularly in public governance, the ability to contest and appeal is essential. This reflects a necessity to correct injustices.
2. **The right to know.** According to human rights, individuals are entitled to explanations of how decisions are made to maintain genuine agency, freedom, and

privacy. Freedom includes the right to receive answers to questions such as, “How am I being tracked? What kinds of inferences are being drawn about me? And how exactly have those inferences been formed?"

3. **A moral obligation exists to understand the consequences of our actions**. As a community, we also bear responsibility for managing risks. This obligation extends, to a reasonable extent, to understanding and anticipating the effects of the technologies we introduce into the world. Thus, claiming “we can’t understand now what it will do” is not a valid argument for launching a system that could cause harm. Instead, it is our moral duty to investigate the potential risks.

These three points can all be summarized as calls for adequate information. Do we know whether and to what extent this algorithmic decision is justified? Do I understand how inferences about me are made? To what extent am I responsible for the actions of the system, and how much should I know about its inner workings to assume that responsibility?

What is transparency?

Transparency can be defined in various ways. Several related concepts are often used synonymously with transparency, such as “explainability” (the AI research area known as “XAI”), “interpretability,” “understandability,” and “black box.”[8]

Transparency is fundamentally a characteristic of an application. It pertains to how well one can understand a system’s inner workings “in theory." It can also refer to providing explanations of algorithmic models and decisions that are accessible to users. This relates to public perception and understanding of how AI operates. Moreover, transparency can be viewed as a broader socio-technical and normative ideal of "openness."

There are many unresolved questions about what defines transparency or explainability, and what level of transparency is sufficient for different stakeholders. The precise interpretation of "transparency" may vary by context. It remains an open scientific question whether multiple forms of transparency exist. Furthermore, transparency can refer to various aspects, whether the aim is to analyze the legal implications of unjust biases or to discuss them in connection with features of machine learning systems.

Transparency as a property of a system

Transparency is a system property that describes how a model works internally. It is further divided into “simulatability” (understanding of the model's functioning), “decomposability” (knowledge of the individual components), and algorithmic transparency (visibility of the algorithms).

What makes a system a “black box”?

Complexity. In contemporary AI systems, the operation of a neural network is encoded in thousands, or even millions, of numerical coefficients. Typically, the system learns its values during the training phase. Because the neural network's operation depends on the complex interactions among these values, it is practically impossible to understand how the network works, even if all the parameters are known.

Difficulty in developing explainable solutions. Even if the AI models support some level of explainability, further development is necessary to enhance the system's explainability. It may

[8] https://www.sciencedirect.com/science/article/pii/S1566253523001148

be challenging to create a user experience that provides careful yet easily understandable explanations.

Risk concerns. Many AI algorithms can be easily fooled if an attacker carefully designs an input that leads the system to malfunction. In a highly transparent system, it may be easier to manipulate it to produce unusual or unwanted results. Consequently, some systems are intentionally created as black boxes. Since many of the most efficient deep learning models are nearly always black-box models, researchers appear to believe that making them fully transparent is highly unlikely. Therefore, the discussion centers on identifying a "sufficient level of transparency." Would it be adequate if algorithms provided people with a disclosure of how decisions were made and specified the smallest change "that can be made to obtain a desirable outcome"?[9] For instance, if an algorithm denies someone a social benefit, it should explain the reason for the denial and outline what actions the individual can take to reverse the decision.

The explanation should indicate, for example, the maximum salary amount that can be approved (input) and how decreasing this amount will influence the decisions made (manipulation of the input). However, the right to understand also applies when the system makes errors. Therefore, it may be necessary to examine the algorithm and identify the factors that led to the system's errors.[10] This cannot be accomplished by simply manipulating the inputs and outputs.

Furthermore, transparency fulfills various other roles in modern discussions about machine learning models. It can be crucial for creating legislation or fostering public trust in AI. To tackle these challenges, the concept of AI transparency is often defined more broadly as "comprehensibility."

Transparency as comprehensibility

The comprehensibility—or understandability—of an algorithm requires explaining it in a way that is sufficiently clear to those affected by the model. One should have a concrete understanding of how or why a particular decision was reached, based on the inputs. However, it is notoriously difficult to translate concepts derived from algorithms into ones that humans can understand. In some countries, legislators have debated whether public authorities should release the algorithms they use for automated decision-making as programming code. However, most people do not know how to interpret programming code. Therefore, it is hard to see how publishing these codes enhances transparency.

Would it be more helpful to publish the exact algorithms? In most cases, publishing the exact algorithms does not bring much transparency, either, especially if you do not have access to the data used to train the model.

Currently, cognitive and computer scientists are developing human-interpretable descriptions of application behavior and reasoning. Their approaches include data visualization tools, interactive interfaces, verbal explanations, and meta-level descriptions of model features. These tools can significantly enhance the accessibility of AI applications. However, much work remains to be done.

The fact that comprehensibility relies on subject and culture-dependent factors complicates the issue further. For instance, the way visualizations are interpreted—and the inferences

[9] https://arxiv.org/pdf/1811.01439

[10] Rusanen & Ylikoski 2017

drawn from them—differ across cultures. Therefore, tech developers should ensure they have a sufficient understanding of the visual language they employ.
Moreover, much is dependent on the degree of data or algorithmic literacy, for example, the knowledge of contemporary technologies. In some cultures, the vocabulary of contemporary technology is more familiar, but in many others they may be completely novel. To increase the understandability, there is clearly a need for significant educational efforts in improving algorithmic literacy – for example, on "computational thinking."[11]. This user literacy will have a direct effect on transparency, as it relates to ordinary users' basic understanding of AI systems. It may actually provide the most efficient and practical way to make the boxes less black for many people.

How to make models more transparent?
The black-box problem in artificial intelligence is not new. Providing transparency for machine learning models is an active area of research. Roughly speaking, there are five main approaches:

1. **Use simpler models**. This, however, often comes at the expense of accuracy for explainability.
2. **Combine simpler and more sophisticated models**. While the sophisticated model enables more complex computations, the simpler model can be used to provide transparency.
3. **Modify inputs to track relevant dependencies between them and outputs**. If manipulating inputs changes the overall model results, these inputs may affect the classification.
4. **Design the models for the user**. This requires using cognitively and psychologically efficient methods and tools for visualizing the model states or directing attention. For example, in computer vision, states in intermediate layers of models can be visualized as features (such as heads, arms, and legs) to provide a comprehensible description for image classification. Researchers have also developed methods to direct "attention" to the parts of the input that matter most. These can be visualized to highlight the parts of an image or a text (so-called "weights") that contribute the most to a particular recommendation.
5. **Follow the latest research**. Much research is ongoing on various aspects of explainable AI, including the socio-cognitive dimensions, and new techniques are being developed. There is a need to translate algorithmic concepts into everyday language. Most people without a background in computer science are not familiar with the basic vocabulary of AI. This directly affects their ability to understand recent developments.

Transparency and the risks of openness
Transparency often signifies a modern, ethical, social, and legal "ideal," representing a normative demand for the responsible use of technology in our societies.[12] It is a reflection of the ideal of "openness", that is framed in terms of "open government," "open data," "open source/code/access", as well as "open science."[13] In this context, transparency considerations

[11] Heintz & al 2016
[12] Koivisto 2016
[13] Larsson 2020

are essential to ensure the equitable distribution of scientific advancements so that the benefits of AI development are accessible to everyone.
Paradoxically, the ideal of openness can also lead to harmful consequences. For instance, the transparency of social media platforms has resulted in various instances of misuse and challenges to democracy. Transparency can create security risks; excessive openness might allow privacy-sensitive data to leak into the wrong hands. Furthermore, the more that is disclosed about algorithms and data, the more harm a malicious actor can inflict. Algorithms are vulnerable to being hacked, and information can make AI more susceptible to intentional attacks. Additionally, entire algorithms can be stolen based purely on their explanations.
In summary, while there is a need to develop more transparent practices for AI, there is also a need to develop practices that can help us avoid abuse. While transparency may help to mitigate ethical issues – such as fairness or accountability – it also creates ethically important risks. Too much openness in the wrong context may defeat the positive development of AI-enabled processes. Taken together, the ideal of full algorithmic transparency should be carefully considered, and we will have to find a balance between security and transparency.

Ethics of AI: Fairness and Non-Discrimination

Should AI Be Fair and Non-Discriminative?

What does fairness mean in relation to artificial intelligence, and how does discrimination manifest through AI systems? These questions are central to one of the core principles of AI ethics: fairness. As artificial intelligence becomes more integrated into decision-making processes in areas such as education, hiring, and finance, it becomes increasingly important to ensure that these systems operate in ways that are just and non-discriminatory.
AI systems have the potential to improve efficiency and expand opportunities. Still, they also carry the risk of reinforcing existing inequalities if they are built on biased data or flawed assumptions. Understanding fairness in AI requires both philosophical insight and practical awareness of how algorithms function in real-world contexts.

What Is Fairness?

Fairness can be understood through real-world scenarios in which algorithmic decision-making affects individuals. Consider a situation where student grades are determined not by exams, but by teacher predictions adjusted by an algorithm based on historical school performance. While the intention of such a system may be to standardize grading and reduce inflation, it can unintentionally disadvantage certain groups, particularly those from lower socio-economic backgrounds.
This example demonstrates that AI systems can appear neutral yet produce unequal outcomes. Fairness is not simply about consistency; it requires examining whether outcomes are just and whether certain groups are disproportionately affected.

Fairness and Bias

Fairness is essential for both social stability and the protection of human dignity. Philosophers such as John Rawls argue that a just society depends on individuals believing they are treated fairly. Similarly, Immanuel Kant emphasizes that all individuals deserve equal moral consideration. When AI systems produce biased outcomes, they risk violating these foundational ethical principles.

At the same time, fairness is inherently complex. Attempts to correct one type of unfairness may unintentionally introduce another. This makes fairness one of the most challenging ethical issues in the development and deployment of AI systems.

Equality and Equity

Two important concepts in discussions of fairness are equality and equity. Equality refers to treating everyone the same, regardless of their circumstances. Equity, on the other hand, involves recognizing differences and providing individuals with what they need to succeed. In the context of AI, these concepts can lead to different design choices. A system that treats all individuals identically may still produce unequal outcomes if it fails to account for structural disadvantages. Conversely, systems designed with equity in mind may intentionally treat individuals differently to achieve more just outcomes.

Types of Justice

Fairness in AI can also be understood through different forms of justice. Distributive justice concerns the fair allocation of resources and opportunities, such as access to jobs or loans. Retributive justice focuses on fairness in punishment, ensuring that penalties are proportionate and unbiased. Compensatory justice addresses how individuals are compensated for harm, which is relevant in cases where AI systems cause harm or disadvantage.

Discrimination and Bias in AI

AI systems can reflect and amplify existing biases in several ways. Language models may learn and reproduce stereotypes present in training data. Hiring algorithms trained on historical data may favor certain groups over others. Credit scoring systems may deny opportunities based on factors that correlate with protected characteristics such as race or gender.

These examples illustrate that AI does not operate in isolation; it inherits patterns from the data and systems that created it. Without careful oversight, AI can perpetuate and even intensify existing inequalities.

What Is Discrimination?

Discrimination can be defined as differential treatment based on membership in a socially significant group that results in harm. Not all differences in treatment are inherently unfair. Still, when those differences are based on characteristics such as race, gender, or socio-economic status, and lead to negative outcomes, they become ethically problematic.

Types of Harm

Discrimination in AI can cause various harms. Allocative harms occur when individuals are denied access to resources or opportunities, such as jobs, loans, or education. Representational harms involve reinforcing stereotypes or misrepresenting groups, thereby shaping societal perceptions and contributing to long-term inequality.

How Bias Enters AI Systems

Bias can enter AI systems through several pathways. One common source is non-representative training data, which does not accurately reflect the diversity of the real world. Another source is label bias, where flawed or biased proxies are used to represent complex

phenomena. Cultural bias in system design can also lead to exclusion, particularly when developers make assumptions that do not account for diverse user experiences.
These sources of bias demonstrate that fairness in AI is not purely a technical issue. It is deeply connected to historical, social, and cultural factors.

Reducing Bias in AI
Efforts to reduce bias in AI include removing sensitive variables from datasets, improving data representation, and regularly auditing systems for fairness. However, these technical solutions are not sufficient on their own. Addressing bias often requires broader changes in organizational practices and societal structures.

Beyond Bias
It is important to recognize that eliminating bias does not automatically make an AI system ethical. Even technically unbiased systems can be used in ways that are harmful or discriminatory. For example, technologies that classify individuals based on personal characteristics can raise serious ethical concerns regardless of their accuracy.
This highlights a key insight in AI ethics: fairness is necessary but not sufficient. Ethical evaluation must consider how systems are used and the broader context in which they operate.

Conclusion
Fairness in AI is a complex and multifaceted issue that requires careful consideration of both technical and ethical factors. As AI continues to shape society, developers, policymakers, and users must work together to ensure that these systems promote justice rather than reinforce inequality.

Should AI Respect And Promote Rights?
What are human rights, and how do they relate to the current ethical guidelines and principles of AI? We'll also examine three rights that are particularly important to AI: the right to privacy, security, and inclusion.
Introduction
During the COVID-19 pandemic, governments struggled to find effective policy-making strategies for safely exiting lockdown. According to epidemiologists, opening society requires efficient tracking, tracing, and monitoring. In many cases, this led to the utilization of various tracing and tracking apps. These apps raised several concerns about privacy and security. Critics saw them as the first steps towards the algorithmic surveillance of citizens.
London authorities decided to try something new. Together with scientists, they developed methods to "capture activity over London" to understand the city's activity levels better. In a project called Odysseus, authorities obtained information on the distribution of activities in London by combining machine learning algorithms, statistical time-series analysis, and image processing.[14] This information about activity in the streets of London can be used to support the safe reopening of streets and public health planning.
In Odysseus, the data is sourced from a wide range of materials. Odysseus combines aggregated, anonymized mobile phone data, anonymized credit card transactions, satellite navigation data, and information from street sensors and traffic cameras. This data is used to

[14] https://www.turing.ac.uk/research/research-projects/project-odysseus-understanding-london-busyness-and-exiting-lockdown

generate counts of vehicles, cyclists, and pedestrians, as well as to indicate the density and effects of social distancing. Great care is taken to ensure data is anonymized so that individuals cannot be identified.
As we take a look at human rights. The right to a safe environment is one of these. Odysseus illustrates how AI can be used to respect and promote the right to safety and a healthy environment. At the same time, the project must take other rights – such as the right to privacy – into account. In London, these concerns were taken seriously. To secure privacy, Odysseus is designed so that all data is anonymized, and individuals cannot be identified from images captured by traffic cameras.
Privacy and security have raised a lot of media attention. They are important, but it is necessary to consider the impact of AI on the full spectrum of fundamental human rights and freedoms as well. How will AI impact the right to education and work, or for a fair trial, to fair and open elections, to freedom of speech, and to assembly and demonstration? And what about special groups, such as children? But first, let us discuss what human rights are.

What are human rights?
Human rights form the foundation of the current ethical guidelines and principles of AI, making them a fundamental component of contemporary AI ethics. Human rights are **universal**: All humans are entitled to them. One does not have to be a particular kind of person or a member of some specific community to have human rights.
Human rights are **norms** that protect all people everywhere from political, legal, and social abuse. They include the following:

- **Civil and political rights**, such as the right to life, liberty, and property, freedom of expression, pursuit of happiness, and equality before the law
- **Social, cultural, and economic rights**, including the right to participate in science and culture, the right to work, and the right to education

The role of human rights is to protect people's ability to form, construe, and pursue their own conceptions of a worthwhile life – it's not just about the ability to live "in liberty, happiness and well-being".

What is a human right?
A human right is a norm that can exist on different levels:

- a shared norm of actual human moralities
- a justified moral norm supported by strong reasons
- a legal right at the national level (where it might be referred to as a "civil" or "constitutional" right)
- a legal right within international law

What is the Universal Declaration of Human Rights?
The Universal Declaration of Human Rights (UDHR) is a document drafted by representatives from diverse legal and cultural backgrounds across all regions of the world.[15] The declaration was proclaimed by the United Nations General Assembly in Paris on December 10, 1948 (General Assembly resolution 217 A) as a common standard of achievement for all peoples and nations.

[15] https://www.un.org/en/about-us/universal-declaration-of-human-rights

Conceptually, human rights are grounded in agency and autonomy[16]. They have an ethical priority: if they compete with other considerations, such as economic wealth, national stability, or other factors, human rights should be prioritized. In the context of AI, this prioritization implies the following requirements:

- AI applications that could clearly violate human rights should not be used
- AI applications that prevent people from enjoying their human rights or actively put them at risk of human rights violations should not be used

However, human rights have context-sensitive properties that allow individuals to prioritize a specific right when needed. Some rights are more fundamental than others. For example, when the right to life conflicts with the right to privacy, the right to privacy will generally be outweighed.

In recent years, privacy and security concerns have dominated discussions of AI and human rights. Emerging combinations of big data analytics, surveillance technologies, and developing biometric recognition methods have recently received significant media and policy attention. Also, the right to equality and inclusion has raised a lot of public discussion. In the next section, we'll take a brief look at these discussions.

Examples of human rights: privacy, security, and inclusion

Privacy

Privacy concerns are raised, for example, by digital records that contain information that can be used to infer sensitive attributes (age, gender, or sexual orientation), preferences, or religious and political views. Biometric data also raises privacy concerns, as it can reveal details of physical and mental health. Often, the real worry is not the data itself but how it can be used to manipulate, affect, or harm a person.

Ethically, privacy is related to personal autonomy and integrity. Following the principles set out by John Locke, a right to control our own personal lives has been seen as central to our autonomy. If that right is taken away, it violates something fundamental about our psychological and moral integrity.

Many have proposed the principle that people should have control over their own data – and that data concerning them should not be used to harm or discriminate against them. According to some, this right to "full control over one's own data" should be recognized as a human right.

But what, exactly, is your "own data"? Is it the raw data, or the collected and analyzed data? If the data is used for secondary purposes, is it still your data? Or, as Wachter and Mittelstadt remark, does the content of inferences that can be drawn from your data belong to your "own data"?[17]

Wachter and Mittelstadt propose that the right to control one's own data should be reformulated as a right to the "right to reasonable inferences". According to them, it is crucial that we can also control the "high-risk inferences" that can be made about us through big data analytics. These inferences are privacy-invasive or reputation-damaging, or have low verifiability (in the sense of being predictive or opinion-based), yet are used for important decisions.

---

[16] Gewirth 1982

[17] Wachter, Sandra, and Brent Mittelstadt. "A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI". LawArXiv, October 12, 2018. Last modified October 12, 2018. osf.io/preprints/lawarxiv/mu2kf_v1.

GDPR

The General Data Protection Regulation (GDPR) is a legal framework that sets guidelines for the collection and processing of personal data from individuals in the European Union.[18]

The GDPR aims to give individuals control over their personal data. Any information that relates to an individual who can be identified directly or indirectly is considered "personal data." This includes names, social security numbers, and email addresses. Location information, biometric data, ethnicity, gender, web cookies, and political or religious beliefs can also qualify as personal data. Pseudonymous data—data that does not directly identify an individual but can be linked to them—can also fall under this definition if it is easy to identify someone from it.

The data subject must provide specific, unambiguous consent to the processing of the data. Consent must be "freely given, specific, informed, and unambiguous." Data subjects can withdraw previously given consent at any time. Children under 13 can only give consent with their parents' permission.

The GDPR recognizes various privacy rights for data subjects, aiming to give individuals greater control over their data. Some of these rights include:

- The right to be informed (a person must be told about the use of their personal data)
- The right of access (it should be explained how someone's personal data is used)
- The right to rectification (a person has the right to be forgotten and the data deleted)
- The right to restrict processing (a person can deny the use of their personal data)

If you process data, you must do so in accordance with the principles of protection and accountability under the GDPR. You must consider these data protection principles when designing any new product or activity. The data protection principles are:

- **Lawfulness, fairness, and transparency**: Processing must be lawful, fair, and transparent to the data subject
- **Purpose limitation**: You must process data for the legitimate purposes specified explicitly to the data subject when you collected it
- **Data minimization**: You should collect and process only as much data as necessary for the purposes specified
- **Accuracy**: You must keep personal data accurate and up to date
- **Storage limitation**: You may only store personally identifying data for as long as necessary for the specified purpose
- **Integrity and confidentiality**: Processing must be done in such a way as to ensure appropriate security, integrity, and confidentiality (for example, by using encryption)
- **Accountability**: The data controller is responsible for being able to demonstrate GDPR compliance with all of these principles

According to GDPR, if you process data, you're also required to handle data securely by implementing "appropriate technical and organizational measures."

How to protect privacy – data anonymization methods

The GDPR allows organizations to collect anonymized data without consent, utilize it for any purpose, and store it indefinitely, provided that they remove all identifiers from the data. Several techniques for data anonymization exist, including:

[18] https://gdpr.eu/what-is-gdpr/

- **Generalization** is a method that deliberately removes some of the data to make it less identifiable. Data can be modified into a set of ranges or a broad area with appropriate boundaries. You can remove the street address while including the town name. In this way, you can eliminate some of the identifiers while retaining a degree of data accuracy.
- **Pseudonymization** is a data management and de-identification method that replaces private identifiers – names, ID-codes – with fake identifiers or pseudonyms, for example, replacing the identifier "Santeri" with "Saara". Pseudonymization preserves statistical accuracy and data integrity. The modified data can be used while still protecting data privacy.
- **Synthetic data** is a method for using artificially created datasets instead of altering the original dataset. The process involves creating statistical models based on patterns found in the original dataset. You can use standard deviations, medians, linear regression, or other statistical techniques to generate the synthetic data.

Data anonymization can be challenging. However, there are methods for "de-anonymization' that attempt to re-identify encrypted or obscured information. De-anonymization, also known as data re-identification, can cross-reference anonymized data with other available information to identify a person, group, or transaction.

Safety and security

The right to safety is a norm that protects individuals from physical, social, and emotional harm, including accidents and malfunctions. Security means safety from malicious and intentional threats.

As a right, safety creates a moral obligation to design our products, laws, and environment in such a way that safety can be maintained even in unconventional circumstances or impairments. In the realm of AI, safety has begun to encompass several distinct conversations, including the following:

1) AI as an existential threat—The discussion around AI as an existential threat adopts a speculative and future-focused perspective regarding artificial intelligence. It centers on the question of what kinds of threats to humanity may arise if AI systems become too complex to control (this type of "superintelligence" scenario is portrayed by thinkers such as Nick Bostrom and Ray Kurzweil). However, the likelihood of a future dominated by superintelligent AI has been questioned by both philosophers and technologists. As it stands, there is no reason to believe that superintelligence will emerge from the development of current algorithmic methods.

2) Safety in AI - The second interpretation of safety in AI addresses the practical issue of designing systems that operate safely and predictably. As AI systems become increasingly integrated into various aspects of life, these systems must be designed to accommodate the world's complexity. A clear example of this is lane guard technology, which employs machine learning to prevent vehicles from drifting out of their lanes. Machine learning researchers have discovered that some lane detection algorithms can be easily misled by road markings, leading cars to veer off the road by following erroneous lane indicators.

One could argue that the right to safety obligates technology producers to account for these scenarios: the fact that the environment was not ideal does not excuse the

system's malfunction. Machine learning researchers refer to this feature as robustness – the system's ability to function predictably under new and unpredictable circumstances.

The ethically and legally significant question is, "What are the acceptable limits to robustness?" It is certainly conceivable that there is a set of circumstances so incredible that even if the system's safety cannot be assured, we can concede that "nobody could have realistically seen that coming." Where this limit is, though, is a difficult problem, and definitely not one that is exclusive to AI or even technology. Nonetheless, the progressive zeal associated with AI future visions has raised wholly new questions about the limits of safety and the taming of environmental uncertainty. An example of this is the discussion of autonomous vehicles.

3) Producing Safety with AI - The third concept of safety and AI that we will explore in this section is the production of safety through the use of AI. Can AI make the world safer? Can AI make the world feel safer? And safer for whom?

Robotization serves as a practical example of this concept. Tasks involving hazardous materials or unsafe environments can be assigned to robots, thus safeguarding the health of human (or animal) workers.

Another way in which certain forms of safety are produced through AI is through automated surveillance. AI-powered surveillance has appeared in various domains: in public spaces, in law enforcement through predictive policing, and in domestic life through products like Amazon's Ring. Although surveillance cameras (CCTV) have long dominated public and semi-public spaces, the ACLU contends that automation introduces a significant qualitative shift in how surveillance operates. But what is so different?

"Imagine a surveillance camera in a typical convenience store in the 1980s. That camera was big and expensive, and connected by a wire running through the wall to a VCR sitting in a back room. There have been significant advances in camera technology in the ensuing decades — in resolution, digitization, storage, and wireless transmission — and cameras have become cheaper and far more prevalent.

"Still, despite all those advances, the social implications of being recorded have not changed: when we enter a store, we typically expect that the presence of cameras won't affect us. We expect our movements to be recorded, and we might feel self-conscious if we notice a camera, particularly if we're doing anything we think might attract attention. However, unless something dramatic happens, we generally understand that the videos in which we appear are unlikely to be scrutinized or monitored."

Constant surveillance creates **"chilling** effects." In other words, knowing that our actions are constantly monitored restricts our genuine freedom to act in the world. Imagine that every time you leave your house, you are followed by two police officers. They never engage with you; they simply stay ten meters behind. You would likely feel uneasy and unable to carry on with your day as usual. In this sense, safety can sometimes conflict with personal freedom and privacy.

Moreover, it remains an open empirical question to what extent AI surveillance truly enhances safety. As illustrated by the chilling effects example, the presence of AI surveillance can contribute to a sense of insecurity. Additionally, it may directly pose risks and cause harm. For instance, AI-powered policing can result in tangible physical

harm due to its predictive nature and enforcement methods. When pervasive and automated surveillance captures even the most minor infractions, it risks making the consequences of policing more harmful than the original offense.

With the disparate levels of policing, disparate methods of enforcement, and disparate levels of surveillance across communities, most clearly along racial dimensions, it becomes clear that AI surveillance creates a different kind of safety (and unsafety) for different people. Again, like before, the value of safety becomes entwined with other ethical values such as justice and non-discrimination.

4) A safe and healthy environment: AI and climate change - Safety also entails the right to a safe and healthy environment. Today, this right is threatened by climate change. The effects of climate change are already apparent; storms, droughts, fires, and flooding have become more common, more frequent, and more devastating. Global ecosystems are shifting, affecting the environment on which our existence relies. The 2018 report on climate change estimated that the world will confront catastrophic consequences unless global greenhouse gas (GHG) emissions are eliminated within thirty years.

AI could be a powerful tool for tackling climate change. It can be used as a resource for monitoring, understanding, and predicting its consequences. AI can accelerate the development of more ecologically sustainable societies. It can be used to design green cities and environmentally friendly transportation, reduce the industry's ecological impact, and develop equipment to help study and maintain ecosystem diversity.

At the same time, many potential problems arise from deploying AI. For instance, innovations aimed at reducing greenhouse gas emissions may inadvertently increase energy consumption and emissions. Given the data and resource-intensive nature of contemporary AI, the technology itself continues to struggle with energy consumption and its carbon footprint. Additionally, one must consider the environmental impact of raw material extraction required for manufacturing AI technologies, which can be significant.

To summarize, safety is involved in AI technologies in multiple ways. These all raise questions about balancing normative values: While calls to make “AI for good” sound promising, in practice, the enactment of rights and normative values in technological systems often collide with the plethora of conflicting interests and deep injustices in the world. When evaluating safety, it is important to evaluate what other rights intersect in practice and ask, “Safety for whom?”

The Limits Of Guidelines

Although ethical principles can shape the development and implementation of ethics-based policy measures and legal norms, empirical studies suggest that guidelines have little impact on the practices surrounding AI development:

“Despite its stated goal, we found no evidence that the ACM code of ethics influences ethical decision-making. Future research is required to identify interventions that do influence decision-making, such as by helping developers identify parallels between their decisions and infamous software news stories.”

Ethics as doing

The question of ethical considerations related to technology is not new. Rather, it has been asked in a very similar form throughout many different iterations of technological development. As an illustration of the historical continuities around these issues, the sociologist Langdon Winner wrote in his 1990 essay "Engineering Ethics and Political Imagination" that a difficulty lies in ethical discussions that center on highly hypothetical and limited "troubling incidents" – without calling into question the broader responsibilities of the entire engineering industry. This issue is similar to those facing the AI ethics project now, 30 years later.

While the troubles may not be new, the contemporary AI ethics project has much more public interest and wider stakeholder participation than engineering ethics has had before. Many more are learning about it, many are hoping to implement AI ethics in their work in some way, and many companies, communities, states, and individuals have stakes in the outcomes. With this new wave of interest, publishing ethical guidelines has become the typical way.

Moving forward with ethics

Moving beyond ethical guidelines, how should AI ethics manifest in the future? What kind of conversations around the ethics of AI should we have, and what kind of activities and ways of doing ethics should be taken into practice? This is a difficult question to answer, but some hints can be found from looking at what is left outside the scope of the AI ethics guidelines we have been discussing so far.

Conclusion: Now it's your turn

Ethical questions regarding AI systems pertain to all stages of the AI system lifecycle, understood here to range from research, design, and development to deployment and use – including maintenance, operation, trade, financing, monitoring and evaluation, validation, end-of-use, disassembly, and termination.

In addition, AI actors can be defined as any actor involved in at least one stage of the AI lifecycle. They can refer to both natural and legal persons, such as researchers, programmers, engineers, data scientists, end users, large technology companies, small and medium enterprises, start-ups, universities, and public entities, among others.

- AI systems raise ethical issues that include, but are not limited to, their impact on decision-making, equality, polarization, and well-being.
- AI systems impact societal sectors such as employment and labor, social interaction, healthcare, education, weaponization, transport, and media.

- AI systems cover topics such as freedom of expression, access to information, privacy, democracy, and discrimination.
- AI systems can also change human experience, challenge human agency, raise concerns over the reliability of information sources, and question the ideal of fundamental dignity.

AI is developing fast. While nobody can say for certain how it will impact our lives, we still can make a difference. As is the case with most emerging technologies, there are real risks. Still, if artificial intelligence is developed and deployed in ethically sustainable ways, AI may bring many positive consequences — not only for individuals, or societies, but for the planet as a whole. The direction of development, however, depends only on us.

“Choice, not chance, determines your destiny.”

-Aristotle