How AI Works

AI Made Easy for Decision Makers

December, 2023

Table of Contents

# About this Course

This is the series of courses in Fred Hutch DaSL’s “AI for Decision Makers” specialization on Coursera.

## 0.1 Specialization Sections

* Introduction
* Course 1: AI Possibilities
* Course 2: Avoiding AI Harm
* Course 3: Establishing AI Infrastructure
* Course 4: AI Policy

## 0.2 Available course formats

This course is available in multiple formats which allows you to take it in the way that best suites your needs. You can take it for certificate which can be for free or fee.

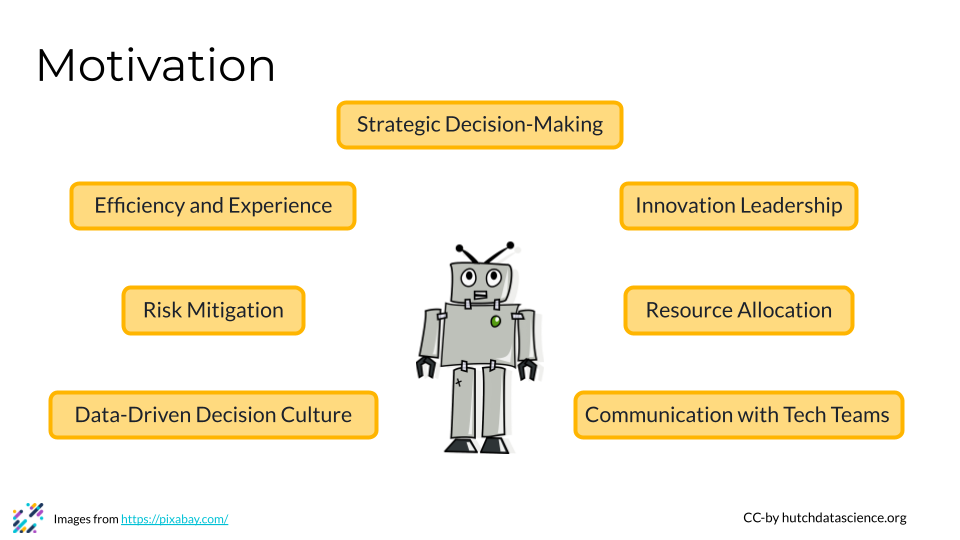
* The material for this course can be viewed without login requirement on this [Bookdown website](LINK%20HERE). This format might be most appropriate for you if you rely on screen-reader technology.
* This course can be taken for [free certification through Leanpub](LINK%20HERE).
* This course can be taken on [Coursera for certification here](LINK%20HERE) (but it is not available for free on Coursera).
* Our courses are open source, you can find the [source material for this course on GitHub](LINK%20HERE).

# 1 VIDEO Summary of This Course

# 2 Introduction

## 2.1 Motivation

How can understanding AI help you be a better leader?



We think understanding AI is essential for executives. It helps today’s leaders make strategic decisions, drive innovation, enhance efficiency, and foster a culture that embraces the transformative power of these technologies. Specifically, AI proficiency can help leaders in the following ways:

1. **Strategic Decision-Making:** Understanding AI and machine learning equips leaders to make informed decisions about integrating these technologies into business strategies, setting their teams up for success when working with AI.
2. **Risk Mitigation:** Familiarity with AI helps leaders assess risks associated with implementing these technologies, ensuring that ethical considerations, data privacy, and potential biases are addressed to mitigate negative consequences. Leaders can also implement more informed policies for their teams.
3. **Efficiency and Experience:** Leaders can explore how AI applications enhance operational efficiency, automate repetitive tasks, and assist employee learning and development, leading to increased productivity and breakthroughs. These improvements can also improve the experience of users or customers your organization serves.
4. **Resource Allocation:** AI resources can be expensive, including in terms of computing resources, subscription services, and/or personnel time. Understanding AI enables leaders to allocate resources effectively, whether in building in-house AI capabilities, partnering with external experts, or investing in AI-driven solutions that align with the organization’s mission.
5. **Innovation Leadership:** Leaders can foster a culture of innovation by understanding the transformative potential of AI. Awareness and knowledge can also enable leaders to identify opportunities for innovation, helping their teams match the rapidly evolving technological landscape.
6. **Data-Driven Decision Culture:** Leaders can promote a data-driven decision-making culture within their organizations, using AI insights to inform strategic planning, understand their teams better, and improve other key business functions.
7. **Communication with Tech Teams:** Executives and managers benefit from understanding AI event if they aren’t building tech, as it helps them effectively communicate with their technical teams. This can mean more effective collaboration and improved alignment between teams or departments.

## 2.2 Target Audience

This specialization is intended for executives, decision-makers, and business leaders across industries, including executives in C-suite positions, managers, and directors. Our goal is for these learners to understand the strategic applications of AI and machine learning in driving innovation, improving operations, creating supportive working environments, and gaining an innovative edge.

We also believe that learning is a life-long process. This specialization is targeted toward those who value continuous learning and want to stay ahead in today’s fast-paced technology landscape.



## 2.3 Curriculum

The course covers…

# 3 Introduction to AI Possibilities

## 3.1 Introduction

This course aims to help decision makers and leaders understand artificial intelligence (AI) at a strategic level. Not everyone will write an AI algorithm, and that is okay! Our rapidly evolving AI landscape means that we need executives and managers who know the essential information to make informed decisions and use AI for good. This course specifically focuses on the essentials of what AI is and what it makes possible, to better harmonize expectations and reality in the workplace.

### 3.1.1 Motivation

This course will help you with your understanding of AI, helping you make strategic decision and cultivate a business environment that embraces the benefits of AI, while understanding its limitations and risks.

### 3.1.2 Target Audience

This course is targeted toward industry and non-profit leaders and decision makers.

### 3.1.3 Curriculum Summary

In this course, we’ll learn about what Artificial intelligence is, and what it isn’t. We’ll also learn the basics of how it works, learn about different types of AI, and set some ground rules for minimizing the harms and maximizing the benefits of AI.

### 3.1.4 Learning Objectives

# 4 VIDEO What Is AI

TODO: Slides here: <https://docs.google.com/presentation/d/1-Mm-Vym3xdtB8xLRHR24jNCLSbNgFHw6N62iJrYe63c/edit#slide=id.g2a162964683_1_0>

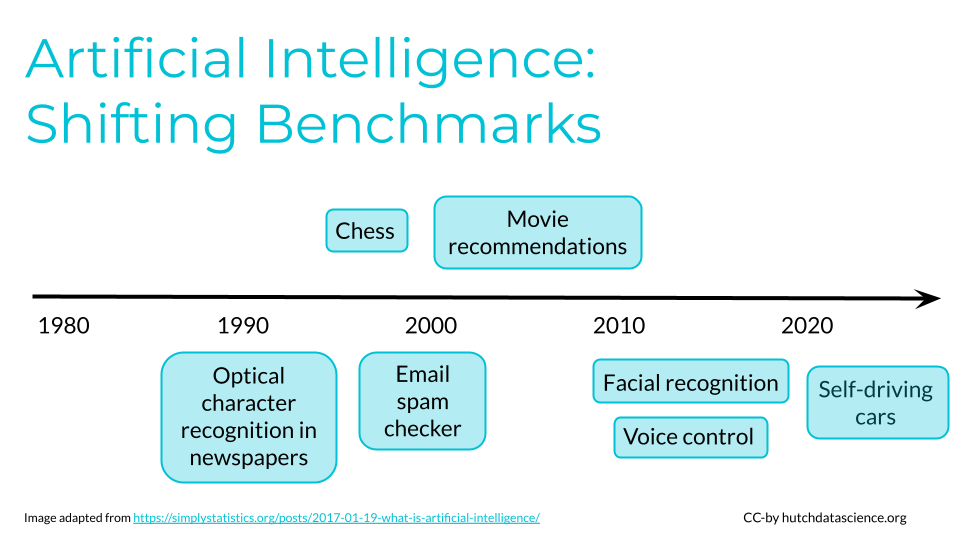
# 5 What Is AI

When discussing artificial intelligence, people often envision humanoid robots, prompting concerns about their ability to outsmart us. The notion of robots passing tests that blur the line between human and machine, often depicted in science fiction, adds to these worries, particularly when considering the potential for AI systems to act in self-interest and make decisions independently.



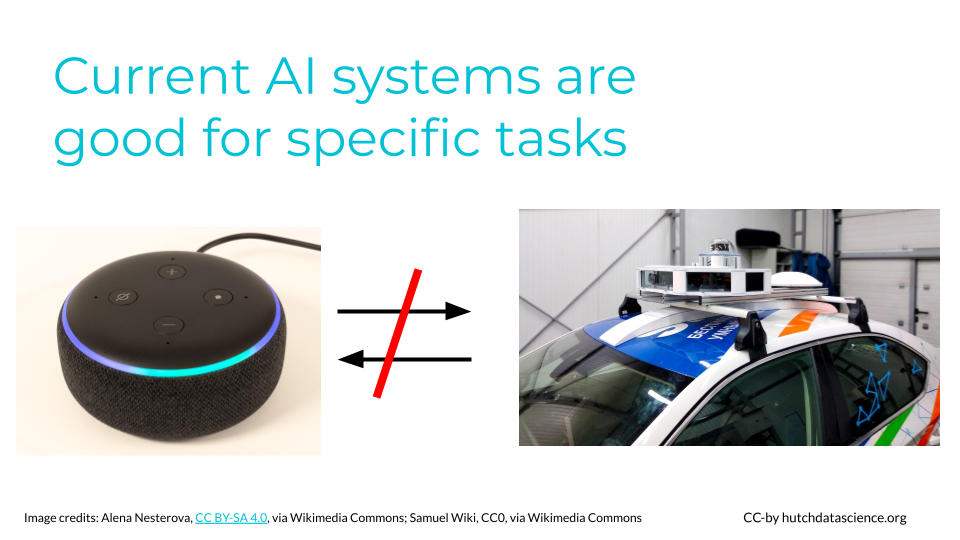
Defining what AI is can be tricky because what experts consider to be AI changes frequently. John McCarthy, one of the leading early figures in AI once said, “As soon as it works, no one calls it artificial intelligence anymore”.

For instance, 20 years ago, the idea of an email spam checker was new. People were surprised that an algorithm could identify junk email accurately, and called it “artificial intelligence”. Since this type of algorithm has become so common, it is no longer called “artificial intelligence”. This transition happened because we no longer think it is surprising that computers can filter spam messages. Because it is not learning something new and surprising, it is no longer considered intelligent.



We often look at human intelligence the same way. For example, many years ago, only a few people knew how to use the internet. These people might have been considered extremely talented and intelligent. Now, the massive growth of online resources and social media mean that fluent internet use is almost required!

Scientists are still working toward computers with full human problem solving and cognition, or **artificial general intelligence**. We aren’t there yet. Currently, artificial intelligence systems are optimized to perform a specific task well, but not for general, multi-purpose tasks. For example, the AI application for recognizing voices can not be directly applied to drive cars, and vice versa. Similarly, a language translation app could not recognize images, and vice versa.

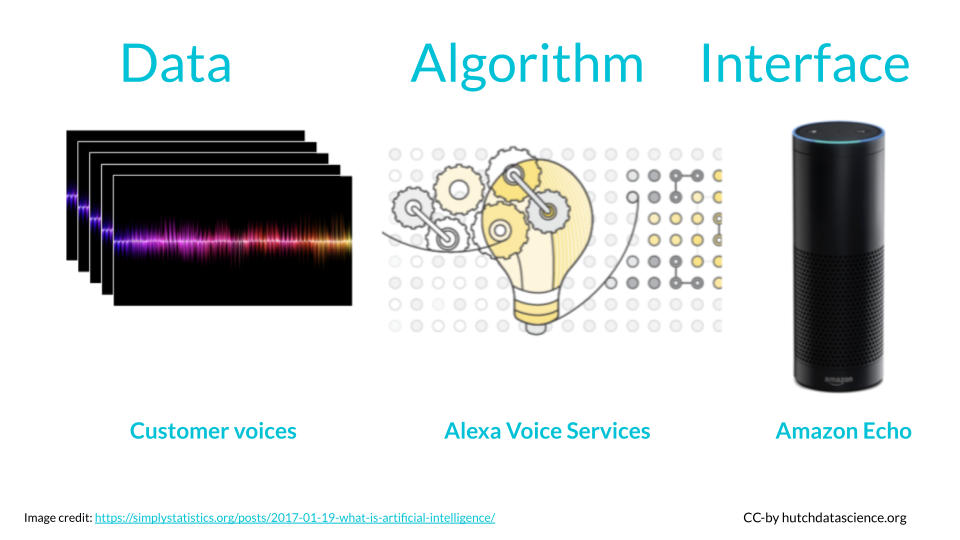


**Artificial General Intelligence (AGI)**: A type of artificial intelligence that can understand, learn, and apply knowledge across a wide range of tasks, similar to the broad cognitive abilities of a human being. It represents the aspiration for machines to have versatile intelligence rather than focusing on specific, narrow domains. Check out **Types of AI** to learn more.

## 5.1 Our AI Definition

Going forward in this course, we define AI as having the following features:

1. **Dataset**: AI needs data examples that can be used to train a statistical or machine learning model to make predictions.
2. **Algorithm**: AI needs an algorithm, or a set of procedures, that can be trained based on the data examples. That way, it can take a new example and execute a human-like task. For instance, the algorithm learns which images feature a cat from pre-labeled images. When given a new image, it decides whether the image has a cat in it.
3. **Interface**: AI needs a physical interface or software for the trained algorithm to receive a data input and execute the human-like task in the real world. For example, you might interface with a chatbot in your web browser.



## 5.2 AI In Practice

The following are case studies that can help us conceptualize AI in the real world.

### 5.2.1 Amazon Recommendations

Amazon’s recommendation engine uses AI algorithms to analyze user behavior and past purchases, providing personalized product recommendations. This enhances the shopping experience, increases customer engagement, and drives sales.

TODO: Text here.

### 5.2.2 Financial Forecasting

TODO: Text here.

## 5.3 What Is and Is Not AI

Let’s look at a few of examples. We can then decide whether or not the examples constitute AI.

### 5.3.1 Smartphones

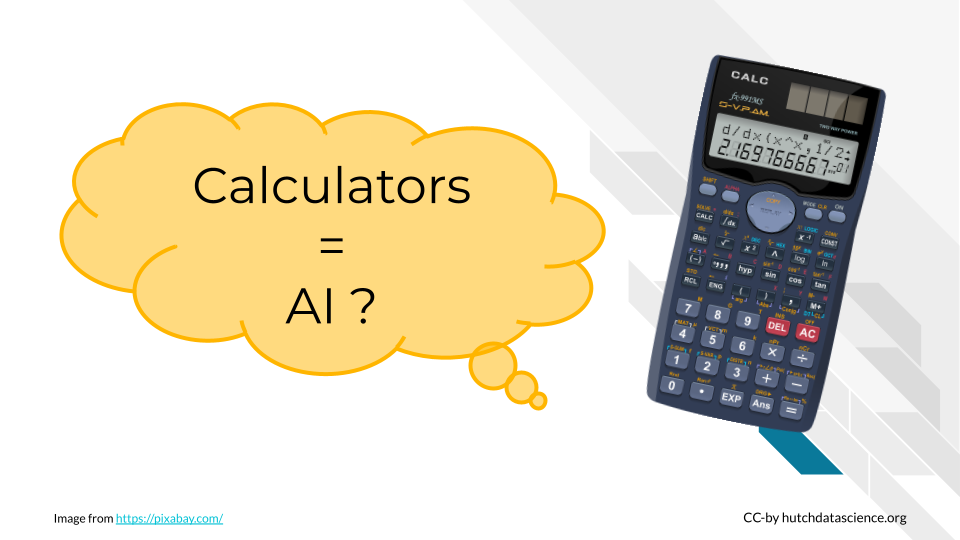


The name “smartphone” implies these devices are making decisions and are powered by AI. Let’s consider our three criteria:

1. **Dataset:** Smartphones do collect a lot of data. For example, they retain your text messages and collect motion tracking information.
2. **Algorithm:** The smartphone as a whole does not usually get trained with this data. However, some features like virtual voice assistants and facial recognition do adapt given your data.
3. **Interface**: Again, some features like voice assistants can be interacted with through the smartphone.

While there are some features on smartphones that are powered by AI models, like virtual voice assistants and facial recognition, the device as a whole isn’t considered AI.

### 5.3.2 Calculators

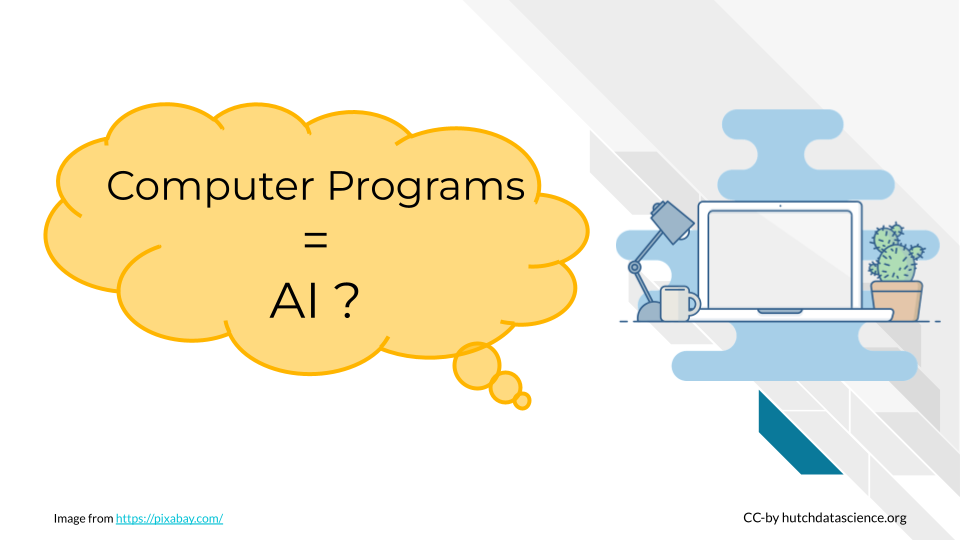


Many of us use basic calculators, as you might find in Microsoft Excel, every day. AI also makes many calculations. Is it just a scaled-up calculator?

1. **Dataset:** Calculators and spreadsheets can store data.
2. **Algorithm:** Calculators do not generally use this data to train algorithms. The procedures that are performed (addition, subtraction, etc.) are almost always predefined. However, some AI-powered assistants are starting to be integrated into software like Excel and Google Sheets.
3. **Interface**: Calculators do meet the criteria for an interface, whether through a physical device or software application.

Traditional calculators are not considered AI, because they can only execute predefined operations.

### 5.3.3 Computer Programs



Like calculators, computers follow set procedures for problem solving and computation. Everyday computers use these procedures to help automate repetitive tasks and save time. However, this isn’t generally considered AI, because the computer’s algorithms aren’t being trained with new data you supply. AI systems exhibit the ability to adapt and handle new inputs for tasks that might be more complicated.

### 5.3.4 DISCUSSION Is It AI

Consider the following examples. Are they examples of AI? Why or why not? Click to expand and see the answer.

**A smartfridge that lets you know when replacement parts are needed**

This is not AI. The computer in the fridge is typically programmed to look for specific signs of wear or time passing. It is not typically trained with new data.

**Speed cameras on the highway**

Speed cameras on highways typically use specialized technology and are not explicitly powered by AI. These cameras are often equipped with radar sensors for measuring vehicle speed between checkpoints.

While the core functionality of speed cameras relies on sensor technology and predetermined speed thresholds, AI elements may be incorporated in some advanced systems. For example, AI could be used to enhance image recognition accuracy for reading license plates. However, the fundamental operation of speed cameras is rooted in sensor-based speed detection, not AI.

**Suggested accounts on Instagram**

This is considered AI. Social media algorithms, like Instagram’s, make recommendations based on user behavior. For example, if you spend a lot of time viewing a page that was recommended, the system interprets that as positive feedback and will make similar recommendations. Typically, these recommendations get better over time as the user generates more user-specific data. You supply data through your behaviors, the algorithm gets trained, and you interact with the suggestions via the app.

# 6 VIDEO How AI Works

# 7 How AI Works

Let’s briefly revisit our definition of AI: it must have data, training via an algorithm, and an interface. How do each of these work? We’ll explore below.

## 7.1 The Data Explosion

Let’s say we’re driving a car or taking public transportation in a city. We might notice a pattern between the amount of traffic on roads, and the time of day. If you commute once at a specific time of day and observe the traffic around you, you have one data point. You can do this a bunch of times and collect more data.

Historically, this is the way data has been collected, and you could manage that data in an Excel Spreadsheet. However, as computer storage has become cheaper and data collection methods have become more sophisticated, our ability to access data has exploded in scale. It’s not hard to imagine that using traffic cameras, dashcams, and car sensors could collect a lot more information than any one person.

Think about how much text information is freely available on the internet! Treating that as input data, AI systems can look for patterns of words that typically go together. For example, you’re much more likely to see the phrase “cancer is a disease” than “cancer is a computer program”.



### 7.1.1 What Is Data

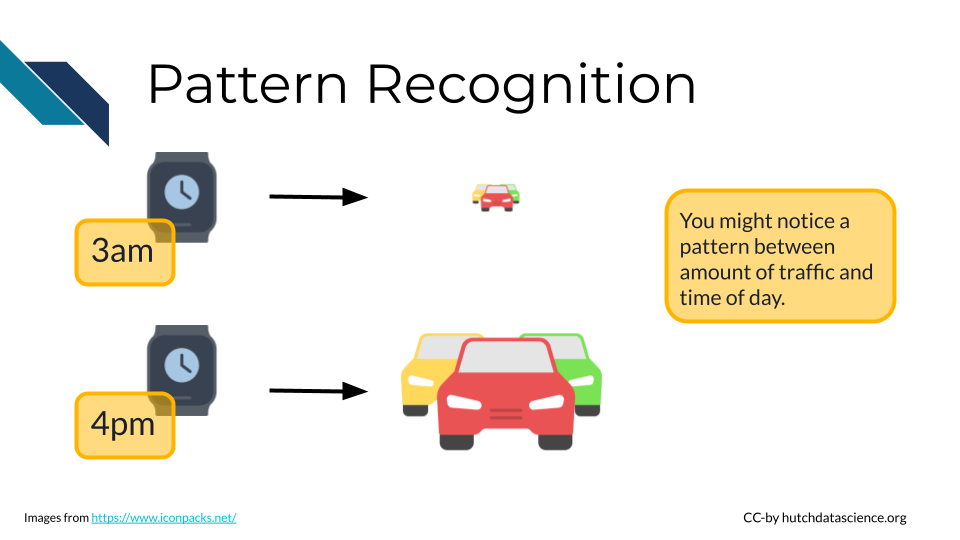
Data comes in many shapes and forms. Data can be **structured**, such as a spreadsheet of times and traffic volume or counts of viral particles in different patients. Data can also be **unstructured**, such as might be found in social media text or genome sequence data.

Other kinds of data can be collected and used to train algorithms. These might include survey data collected directly from consumers, medical data collected in a healthcare setting, purchase or transaction tracking, and online tracking of your time on certain web pages ([Cote 2022](#ref-Cote2022)).

It is **essential** that you and your team think critically about data sources. Many companies releasing generative AI systems have come under fire for training these systems on data that doesn’t belong to them ([Walsh 2023](#ref-Walsh2023)). Individual people also have a right to data privacy. No personal data should be used without permission, even if that data could be interesting or useful.

## 7.2 Machines Can Learn Like Us

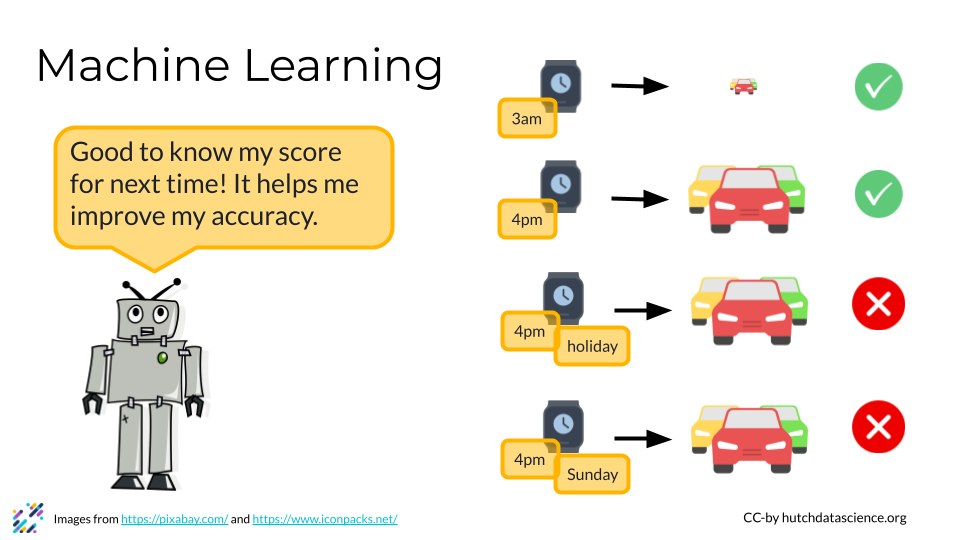
Human beings are powerhouses when it comes to pattern recognition and processing ([Mattson 2014](#ref-Mattson2014)). We are constantly observing the world around us, collecting data to learn and make decisions. For example, we might notice a pattern between the amount of traffic on roads in a city, and the time of day.



Much like the human brain, machine learning detects patterns within data. **Machine learning** is at the heart of artificial intelligence, allowing computers to learn and make predictions. In more complex machine learning, computers make millions of calculations, mastering the mapping of inputs (observations) to outputs (predictions). This process mirrors how humans learn through experience.

**Machine Learning**: Machine learning is a way for computers to learn from examples and improve their performance over time, resembling how humans learn from experience.

A machine learning system refines its understanding by continuously updating its parameters based on the feedback received from the provided data. For example, our system might be guessing traffic by time of day, but also judging its accuracy while accounting for other factors, such as whether or not it was a work day, if some workers are on holiday, or how many people live in the city.



The rise of machine learning has been propelled by our ability to collect vast amounts of data and sophisticated types of AI and computing power.

# 8 VIDEO Different Types of AI

# 9 Types of AI

How they work..

# 10 VIDEO Real Life Possibilities

What type of AI for specific possibilities - case studies

# 11 What Is Possible

What is possible with AI? What’s still fantasy?

# 12 VIDEO What Is Possible

# 13 VIDEO What Is NOT Possible

# 14 Ground Rules for AI

Ground rules - don’t do bad things with AI!

# 15 VIDEO Knowing the Ground Rules

# 16 Introduction to Avoiding AI Harm

This course aims to help you recognize some of the potential consequences of using or developing AI tools.

Some of this content was adapted from our course on [AI for Efficient Programming](https://hutchdatascience.org/AI_for_Efficient_Programming/ethics-of-using-ai.html). If you intend to use AI for writing code, we recommend that you review this content for a deeper dive into ethics specifically for writing code with generative AI.

## 16.1 Motivation

The use of artificial intelligence (AI) and in particular, generative AI, has raised a number of ethical concerns. We will highlight several current concerns, however please be aware that this is a dynamic field and the possible implications of this technology is continuing to develop. It is critical that we as a society continue to evaluate and predict what the consequences of the use of AI will be, so that we can mitigate harmful effects.



## 16.2 Target Audience

This course is intended for leaders who might make decisions about AI at nonprofits, in industry, or academia.

## 16.3 Curriculum

This course provides a brief introduction about ethical concepts to be aware of when making decisions about AI. It will cover:

* A proposed process for Ethical AI use and development
* Possible societal impacts of AI
* Concerns to be aware of for AI algorithms
* Strategies to adhere to AI codes of ethics
* Concepts for consent with AI
* IDARE principles (Inclusion, Diversity, Anti-Racism, and Equity) with AI

## 16.4 Learning Objectives

In this chapter we will demonstrate how to:

* Explain a possible process for ethical AI use and development
* Describe key ethical concerns for using AI tools
* Identify possible mitigation strategies for these major concerns
* Explain the potential benefits of being transparent about the use of AI tools
* Discuss why human contributions are still important and necessary
* Recognize strategies for using AI tools more responsibly

# 17 Ethical process

The concepts for ethical AI use are still highly debated as this is a rapidly evolving field.

However, here is a proposed framework for using AI more ethically.

# 18 Societal Impact

AI has the potential to dramatically influence society. It is our responsibility to be proactive in thinking about what uses and impacts we consider to be useful and appropriate and those we consider harmful and in appropriate.

There are a few current major codes of ethics for AI. As this is an emerging technology, more codes will be developed and updated as the technology evolves.

* United States [AI Bill of Rights](https://www.whitehouse.gov/ostp/ai-bill-of-rights/)
* European Commission [Ethics Guidelines for trustworthy AI](https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai)
* The Institute of Electrical and Electronics Engineers (IEEE) [Ethically Aligned Design Version 2](https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf)

In this chapter we will discuss the some of the major ethical considerations in terms of possible societal consequences for the use or development of AI tools:

1. **Intentional and Inadvertent Harm** - Data and technology intended to serve one purpose may be reused by others for unintended purposes. How do we prevent intentional harm?
2. **Replacing Humans and Human autonomy** - AI tools can help humans, but they are not a replacement. Humans are still much better at generalizing their knowledge to other contexts and human oversight is required.
3. **Inappropriate Use** - There are situations in which using AI might not be appropriate now or in the future.
4. **Bias Perpetuation and Disparities** - AI models are built on data and code that were created by biased humans, thus bias can be further perpetuated by using AI tools. In some cases Bias can even be exaggerated. This combined with differences in access may exacerbate disparities.
5. **Security and Privacy Issues** - Data for AI systems should be collected in an ethical manner that is mindful of the rights of the individuals the data comes from. Data around usage of those tools should also be collected in an ethical manner. Commercial tool usage with proprietary or private data, code, text, images or other files may result in leaked data not only to the developers of the commercial tool, but potentially also to other users.
6. **Climate Impact** - As we continue to use more and more data and computing power, we need to be ever more mindful of how we generate the electricity to store and perform our computations.

Note that this is an incomplete list; additional ethical concerns will become apparent as we continue to use these new technologies. We highly suggest that users of these tools **careful to learn more about the specific tools they are interested in** and to be **transparent** about the use of these tools, so that as new ethical issues emerge, we will be better prepared to understand the implications.

## 18.1 Intentional and Inadvertent Harm

AI tools need to be developed with safeguards and continually audited to ensure that the AI system is not responsive to harmful requests by users. With additional usage and updates, AI tools can adapt and thus continual auditing is required.

Of course using AI to help you perform a harmful action would result in intentional harm. This may sound like an obvious and easy issue to avoid, at least by those with good intent. However, the consequences may be much further reaching than might be first anticipated.

Perhaps you or your company develop an AI tool that helps to identify individuals that might especially benefit from a product or service that you offer. This in and of itself is likely not harmful. However, the data you have used, the data that you may have collected, and the tool that you have created, all could be used for other malicious reasons, such as targeting specific groups of people for advertisements when they are vulnerable.

Therefore it is critical that we be considerate of the downstream consequences of what we create and what might happen if that technology or data was used for other purposes.

### 18.1.1 Tips for avoiding inadvertent harm

* Consider how the content, data, or newly developed AI tool might be used by others.
* Continually audit any new AI tools.

## 18.2 Replacing Humans and Human Autonomy

While AI systems are useful, they **do not replace human strengths**. Humans remain far superior at generalizing concepts to new contexts.

Computer science is a field that has historically lacked diversity. It is critical that we support diverse new learners of computer science, as we will continue to need human involvement in the development and use of AI tools. This can help to ensure that more diverse perspectives are accounted for in our understanding of how these tools should be used responsibly.

### 18.2.1 Tips for supporting human contributions

* Avoid thinking that content by AI tools must be better than that created by humans, as this is not true.
* Recall that humans wrote the code to create these AI tools and that the data used to train these AI tools also came from humans. Many of the large commercial AI tools were trained on websites and other content from the internet.
* Be transparent where possible about when you do or do not use AI tools, give credit to the humans involved as much as possible.

A new term in the medical field called [AI paternalism](https://www.technologyreview.com/2023/04/21/1071921/ai-is-infiltrating-health-care-we-shouldnt-let-it-make-decisions/) describes the concept that doctors (and others) may trust AI over their own judgment or the experiences of the patients they treat. This has already been shown to be a problem with earlier AI systems intended to help distinguish patient groups. Not all humans will necessarily fit the expectations of the AI model if it is not very good at predicting edge cases ([Hamzelou n.d.](#ref-AI_paternalism)). Therefore, in all fields it is important for us to not forget our value as humans in our understanding of the world.

## 18.3 Inappropriate Use

There are situations in which we may, as a society, not want an automated response. There may even be situations in which we do not want to bias our own human judgment by that of an AI system. There may be other situations where the efficiency of AI may also be considered inappropriate. While many of these topics are still under debate and AI technology continues to improve, we challenge the readers to consider such cases given what is currently possible and what may be possible in the future.

Some reasons why AI may not be appropriate for certain situation include:

* Despite the common misconception that AI systems have clearer judgment than humans, they are in fact typically just as prone to bias and sometimes even exacerbate bias (Pethig and Kroenung ([2023](#ref-pethig_biased_2023))). There are some very mindful researchers working on these issues in specific contexts and making progress where AI may actually improve on human judgment, but generally speaking AI systems are currently typically biased and reflective of human judgment but in a more limited manner based on the context in which they have been trained.
* AI systems can behave in unexpected ways (Gichoya et al. ([2022](#ref-gichoya_ai_2022))).
* Humans are still better than AI at generalizing what they learn for new contexts.
* Humans can better understand the consequences of discussions from a humanity standpoint.

Some examples where it may be considered inappropriate for AI systems to be used include:

* In the justice system to determine if someone is guilty of a crime or to determine the punishment of someone found guilty of a crime.
* It may be considered inappropriate for AI systems to be used in certain warfare circumstances.

### 18.3.1 Tips for avoiding inappropriate uses

* Stay up-to-date on current practices and standards for your field, as well as up-to-date on the news for how others have experienced their use of AI.
* Stay involved in discussions about appropriate uses for AI, particularly for policy.
* Begin using AI slowly and iteratively to allow time to determine the appropriateness of the use. Some issues will only be discovered after some experience.
* Involve a diverse group of individuals in discussions of intended uses to better account for a variety of perspectives.
* Seek outside expert opinion whenever you are unsure about your AI use plans.
* Consider AI alternatives if something doesn’t feel right.

## 18.4 Bias Perpetuation and Disparities

One of the biggest concerns is the potential for AI to further perpetuate bias. AI systems are trained on data created by humans. If this data used to train the system is biased (and this includes existing code that may be written in a biased manner), the resulting content from the AI tools could also be biased. This could lead to discrimination, abuse, or neglect for certain groups of people, such as those with certain ethnic or cultural backgrounds, genders, ages, sexuality, capabilities, religions or other group affiliations.

It is well known that data and code are often biased ([Belenguer 2022](#ref-belenguer_ai_2022)). The resulting output of AI tools should be evaluated for bias and modified where needed. Please be aware that because bias is intrinsic, it may be difficult to identify issues. Therefore, people with specialized training to recognize bias should be consulted. It is also vital that evaluations be made throughout the software development process of new AI tools to check for and consider potential perpetuation of bias.

Because of differences in access to technology, disparities may be further exacerbated by the usage of AI tools. Consideration and support for under-served populations will be even more necessary. For example tools that only work well on individuals with light skin, will lead to further challenges to some individuals.

Developing and scaling-up artificial intelligence-based innovations for use in low- and middle-income countries will thus require deliberate efforts to generate locally representative training data (Paul and Schaefer ([2020](#ref-paul_safeguards_2020))).

### 18.4.1 Tips for avoiding bias

* Be aware of the biases in the data that is used to train AI systems.
* Check for possible biases within data used to train new AI tools.
  + Are there harmful data values? Examples could include discriminatory and false associations.
  + Are the data adequately inclusive? Examples could include a lack of data about certain ethnic or gender groups or disabled individuals, which could result in code that does not adequately consider these groups, ignores them all together, or makes false associations.
  + Are the data of high enough quality? Examples could include data that is false about certain individuals.
* Evaluate the code for new AI tools for biases as it is developed. Check if any of the criteria for weighting certain data values over others are rooted in bias.
* Consider the possible outcomes of the use of content created by AI tools. Consider if the content could possibly be used in a manner that will result in discrimination.

See Belenguer ([2022](#ref-belenguer_ai_2022)) for more guidance. We also encourage you to check out the following video for a classic example of bias in AI:

For further details check out this [course](https://www.coursera.org/learn/algorithmic-fairness) on Coursera about building fair algorithms. We will also describe more in the next section.

## 18.5 Security and Privacy issues

Security and privacy are a major concern for AI usage. Here we discuss a few aspects related to this.

### 18.5.1 Use the right tool for the job

There are two kinds of commercial AI tools (([**nigro\_ai\_nodate?**](#ref-nigro_ai_nodate))):

* those that are designed for public general use
* those that are disgned private use with sensitive data

Public commercial AI tools are often **not designed to protect users from unknowingly submitting prompts that include propriety are private information**. Different AI tools have different practices in terms of how they do or do not collect data about the prompts that people submit. They also have different practices in terms of if they reuse information from prompts to other users. Note that the AI system itself may not be trained on responses for how prompt data is collected or not. So asking the AI system may not give accurate answers.

Thus if users of public AI tools, such as ChatGPT submit prompts that include propriety or private information, they **run the risk of that information being viewable not only by the developers/maintainers of the AI tool used, but also by other users** who use that same AI tool.

### 18.5.2 AI can have security blind spots

Furthermore, AI tools are not always trained in a way that is particularly conscious of data security. If for example, code is written using these tools by users who are less familiar with coding security concerns, protected data or important passwords may be leaked within the code itself. AI systems may also utilize data that was actually intended to be private.

### 18.5.3 Data source issues

It is also important to consider what data the responses that you get from a commercial AI tool might actually be using. Are these datasets from people who consented to their data being used in this manner? If you are generating your own tools, did people consent for their data to be used as you intend?

Data privacy is a major issue all on it’s own:

98% of Americans still feel they should have more control over the sharing of their data (Pearce ([2021](#ref-pearce_beware_2021)))

It is important to follow legal and ethical guidance around the collection of data and to use tools that also abide by these guidelines.

### 18.5.4 Tips for reducing security and privacy issues

* Check that no sensitive data, such as Personal Identifiable Information (PII) or propriety information becomes public through prompts to commercial AI systems.
* Consider purchasing a license for a private AI system if needed or create your own if you wish to work with sensitive data (seek expert guidance to determine if the AI systems are secure enough).
* Promote for regulation of AI tools by voting for standards where possible.
* Ask AI tools for help with security when using commercial tools, but to not rely on them alone. In some cases, commercial AI tools will even provide little guidance about who developed the tool and what data it was trained on, regardless of what happens to the prompts and if they are collected and maintained in a secure way.
* Consult with an expert about data security if you want to design or use a AI tool that will regularly use private or propriety data.
* Development of new tools should involve

Are there any possible data security or privacy issues associated with the plan you proposed?

## 18.6 Climate Impact

The data storage and computing resources needed for AI could exacerbate climate challenges.

## 18.7 Tips for reducing climate impact

Thoughtful planning to efficiently modify existing models as opposed to unnecessarily creating new models from scratch is needed. Solutions such as [federated learning](https://research.ibm.com/blog/what-is-federated-learning), where AI models are iteratively trained in multiple locations using data at those locations, instead of collectively sharing the data to create more massive datasets can help [reduce](https://www.cam.ac.uk/research/news/can-federated-learning-save-the-world.) the required resources and also help preserve data privacy and security.

## 18.8 Summary

Here is a summary of all the tips we suggested:

* Be mindful of how content created with AI or AI tools may be used for unintended purposes.
* Be aware that humans are still better at generalizing concepts to other contexts.
* Always have expert humans review content created by AI and value human contributions and thoughts.
* Carefully consider if an AI solution is appropriate for your context.
* Be aware that AI systems are biased and their responses are likely biased. Any content generated by an AI system should be evaluated for potential bias.
* Be aware that AI systems may behave in unexpected ways. Implement new AI solutions slowly to account for the unexpected. Test those systems and try to better understand how they work in different contexts.
* Be aware of the security and privacy concerns for AI, be sure to use the right tool for the job and train those at your institute appropriately.
* Consider the climate impact of your AI usage and proceed in a manner makes efficient use of resources.

Overall, we hope that awareness of these concerns and the tips we shared will help us all use AI tools more responsibly. We recognize however, that as this is emerging technology and more ethical issues will emerge as we continue to use these tools in new ways. Staying up-to-date on the current ethical considerations will also help us all continue to use AI responsibly.

# 19 Algorithm considerations

In this chapter we will discuss the some of the major ethical considerations regarding the algorithms underlying AI tools. We will provide some tips for how to deal with these issues that may be useful for creating AI guidelines at your institution.

1. **Toxic Responses** - Currently it is not clear how well generative AI models restrict harmful responses in terms of ideas, code, text, etc.
2. **Lack of Interpretability** - When complicated algorithms are used within AI systems, it can be unclear how it came up with a decision. In many circumstances it is necessary to understand how the AI system works to know how to proceed.
3. **Misinformation and Faulty Responses** - Fake or manipulated data used to help design algorithms could be believed to be correct and this could be further propagated. Text, code, etc. provided to users may not be correct or optimal for a given situation, and may have at times severe downstream consequences.
4. **Copyright Violations** - AI model responses are often not transparent about using code, text, images and other data types that may violate copyright.

Note that this is an incomplete list; additional ethical concerns will become apparent as we continue to use these new technologies. We highly suggest that users of these tools **careful to learn more about the specific tools they are interested in** and to be **transparent** about the use of these tools, so that as new ethical issues emerge, we will be better prepared to understand the implications.

## 19.1 Harmful or Toxic Responses

Another major concern is the use of AI to generate malicious content or that AI itself may accidentally create harmful responses. For instance, AI could start suggesting the creation of code that spreads malware or hacks into computer systems. Another issue is what is called [“toxicity”](https://towardsdatascience.com/toxicity-in-ai-text-generation-9e9d9646e68f), which refers to disrespectful, rude, or hateful responses (Nikulski ([2021](#ref-nikulski_toxicity_2021))). These responses can have very negative consequences for users. Ultimately both issues could cause severe damage to individuals and organizations, including data breaches and financial losses. AI systems need to be designed with safeguards to avoid harmful responses, to test for such responses, and to ensure that the system is not infiltrated by additional possibly harmful parties.

### 19.1.1 Tips for avoiding the creation of harmful content

* Be careful about what commercial tools you employ, they should be transparent about what they do to avoid harm.
* If designing a system, ensure that best practices are employed to avoid harmful responses. This should be done during the design process and should the system should also be regularly evaluated. Some development systems such as [Amazon Bedrock](https://aws.amazon.com/blogs/aws/evaluate-compare-and-select-the-best-foundation-models-for-your-use-case-in-amazon-bedrock-preview/) have tools for evaluating [toxicity](https://towardsdatascience.com/toxicity-in-ai-text-generation-9e9d9646e68f) to test for harmful responses. Although such systems can be helpful to automatically test, evaluation should also be done directly by humans.
* Be careful about the context in which you might have people use AI - will they know how to use it responsibly?
* Be careful about what content you share publicly, as it could be used for malicious purposes.
* Consider how the content might be used by others.
* Ask the AI tools to help you, but do not rely on them alone.

What are the possible downstream uses of this content?

What are some possible negative consequences of using this content?

## 19.2 Lack of Interpretability

There is risk in using AI tools, that we may encounter situations where it is unclear why the AI system came to a particular result. AI systems that use more complicated algorithms can make it difficult to trace back the decision process of the algorithm. Using content created or modified by AI, could make it difficult for others to understand if the content is adequate or appropriate, or to identify and fix any issues that may arise.

This could result in negative consequences, such as for example reliance on a system that distinguishes consumers or patients based on an arbitrary factor that is actually not consequential. Decisions based on AI responses therefore need to be made extra carefully and with clarity about why the AI system may be indicating various trends or predictions.

### 19.2.1 Tips for avoiding a lack of interpretability

* Content should be reviewed by those experienced in the given field.
* Ask AI tools to help you understand the how it got to the response that it did, but get expert assistance where needed.

Can you explain how you generated this response?

## 19.3 Misinformation and Faulty Responses

AI tools use data that may contain false or incorrect information and may therefore respond with content that is also false or incorrect.

This is due to number of reasons:

* AI tools may “hallucinate” fake response based on artifacts of the algorithm
* AI tools may be trained on data that is out-of-date
* AI tools may be trained on data that has fake or incorrect information
* AI tools are not necessarily trained for every intended use and may therefore may not reflect best practices for a given task or field

AI tools may also report data as if it is real, when it is in fact not real. For example, currently at the time of the writing of this course, ChatGPT will report citations with links that are not always correct. Furthermore, AI models can “hallucinate” incorrect responses based on artifacts of the algorithm underneath the tool. These responses are essentially made up by the tool. It is difficult to know when a tool is hallucinating especially if it is a tool that you did not create, therefore it is important to review and check responses from AI tools. There is also a risk that content written with AI tools, may be incorrect or inappropriate for the given context of intended use, or they may not reflect best practices for a given context or field. The tools are limited to the data they were trained on, which may not reflect your intended use.

It is also important to remember that content generated by AI tools is not necessarily better than content written by humans. It requires just as much, if not more review.

### 19.3.1 Tips for reducing misinformation & faulty responses

* Be aware that some AI tools currently make up false information based on artifacts of the algorithm called hallucinations or based on false information in the training data.
* Do not assume that the content generated by AI is real or correct.
* Realize that AI is only as good or up-to-date as what it was trained on, the content may be generated using out-of-date data. Look up responses to ensure it is up-to-date.
* In many cases **utilizing multiple AI tools** can help you to **cross-check** the responses (however be careful about the privacy of each tool if you use any private or propriety data in your prompts!).
* Ask the AI tools for extra information about if there are any potential limitations or weaknesses in the responses, but keep in mind that the tool may not be aware of issues and therefore human review is required. The information provided by the tool can however be a helpful starting point.

Are there any limitations associated with this response?

What assumptions were made in creating this content?

[Stack Overflow](https://stackoverflow.com/), a popular community-based website where programmers help one another, has (at the time of writing this) temporarily [banned](https://meta.stackoverflow.com/questions/421831/temporary-policy-chatgpt-is-banned) users from answering questions with AI-generated code. This is because users were posting incorrect answers to questions. It is important to follow this policy (as you may face removal from the community). This policy goes to show that you really need to check the code that you get from AI models. While they are currently helpful tools, they do not know everything.

## 19.4 Lack of Interpretability

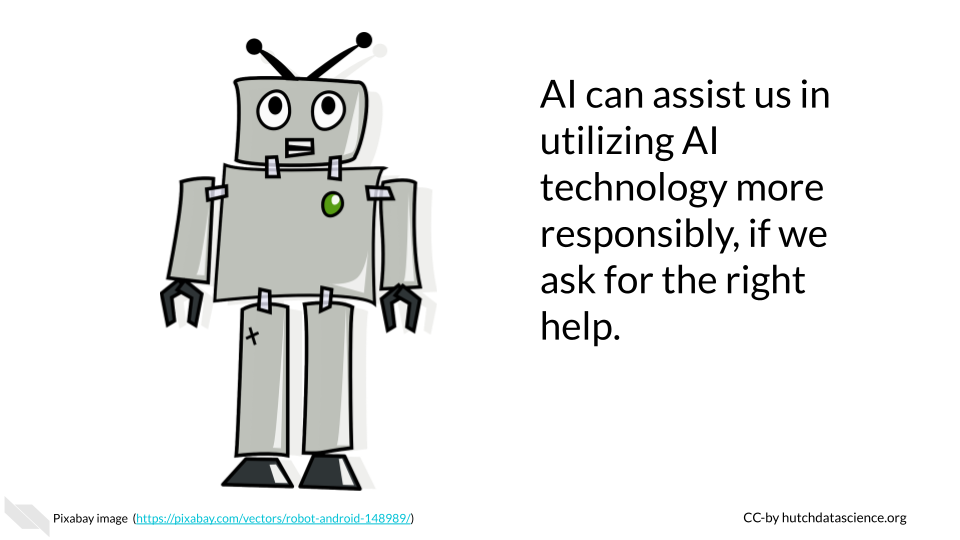
There is risk in using AI tools, that we may encounter situations where it is unclear why the AI system came to a particular result. AI systems that use more complicated algorithms can make it difficult to trace back the decision process of the algorithm. Using content created or modified by AI, could make it difficult for others to understand if the content is adequate or appropriate, or to identify and fix any issues that may arise.

This could result in negative consequences, such as for example reliance on a system that distinguishes consumers or patients based on an arbitrary factor that is actually not consequential. Decisions based on AI responses therefore need to be made extra carefully and with clarity about why the AI system may be indicating various trends or predictions.

### 19.4.1 Tips for avoiding a lack of interpretability

* Content should be reviewed by those experienced in the given field.
* Ask AI tools to help you understand the how it got to the response that it did, but get expert assistance where needed.

Can you explain how you generated this response?



## 19.5 Misinformation and Faulty Responses

AI tools use data that may contain false or incorrect information and may therefore respond with content that is also false or incorrect.

This is due to number of reasons:

* AI tools may “hallucinate” fake response based on artifacts of the algorithm
* AI tools may be trained on data that is out-of-date
* AI tools may be trained on data that has fake or incorrect information
* AI tools are not necessarily trained for every intended use and may therefore may not reflect best practices for a given task or field

AI tools may also report data as if it is real, when it is in fact not real. For example, currently at the time of the writing of this course, ChatGPT will report citations with links that are not always correct. Furthermore, AI models can “hallucinate” incorrect responses based on artifacts of the algorithm underneath the tool. These responses are essentially made up by the tool. It is difficult to know when a tool is hallucinating especially if it is a tool that you did not create, therefore it is important to review and check responses from AI tools. There is also a risk that content written with AI tools, may be incorrect or inappropriate for the given context of intended use, or they may not reflect best practices for a given context or field. The tools are limited to the data they were trained on, which may not reflect your intended use.

It is also important to remember that content generated by AI tools is not necessarily better than content written by humans. It requires just as much, if not more review.

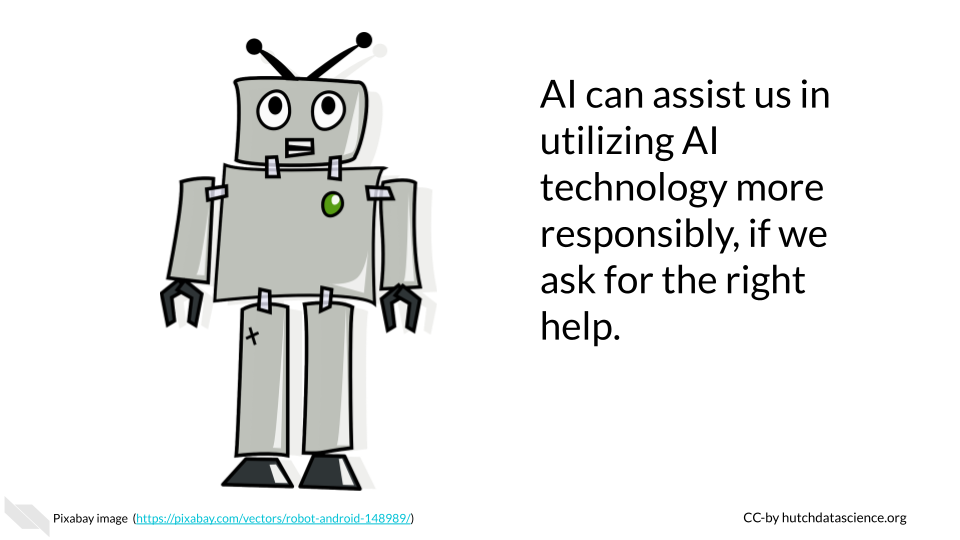
### 19.5.1 Tips for reducing misinformation & faulty responses

* Be aware that some AI tools currently make up false information based on artifacts of the algorithm called hallucinations or based on false information in the training data.
* Do not assume that the content generated by AI is real or correct.
* Realize that AI is only as good or up-to-date as what it was trained on, the content may be generated using out-of-date data. Look up responses to ensure it is up-to-date.
* In many cases **utilizing multiple AI tools** can help you to **cross-check** the responses (however be careful about the privacy of each tool if you use any private or propriety data in your prompts!).
* Ask the AI tools for extra information about if there are any potential limitations or weaknesses in the responses, but keep in mind that the tool may not be aware of issues and therefore human review is required. The information provided by the tool can however be a helpful starting point.

Are there any limitations associated with this response?

What assumptions were made in creating this content?

[Stack Overflow](https://stackoverflow.com/), a popular community-based website where programmers help one another, has (at the time of writing this) temporarily [banned](https://meta.stackoverflow.com/questions/421831/temporary-policy-chatgpt-is-banned) users from answering questions with AI-generated code. This is because users were posting incorrect answers to questions. It is important to follow this policy (as you may face removal from the community). This policy goes to show that you really need to check the code that you get from AI models. While they are currently helpful tools, they do not know everything.



## 19.6 Security and Privacy issues

Commercial AI tools are often not designed to protect users from unknowingly submitting prompts that include propriety are private information. Different AI tools have different practices in terms of how they do or do not collect data about the prompts that people submit. They also have different practices in terms of if they reuse information from prompts to other users. Thus if users submit prompts that include propriety or private information, they run the risk of that information being viewable not only by the developers/maintainers of the AI tool used, but also by other users who use that same AI tool. Note that the AI system itself may not be trained on responses for how prompt data is collected or not.

Furthermore, AI tools are not always trained in a way that is particularly conscious of data security. If for example, code is written using these tools by users who are less familiar with coding security concerns, protected data or important passwords may be leaked within the code itself. AI systems may also utilize data that was actually intended to be private.

It is also important to consider what data your the responses that you get from an AI tool might actually be using.

## 19.7 Violating Copyright

When AI systems are trained on data, they may also learn and incorporate copyrighted information. This means that AI-generated content could potentially infringe on the copyright of the original author. For example, if an AI system is trained on a code written by a human programmer, the AI system could generate code that is identical to or similar to the code from that author. If the AI system then uses this code without permission from the original author, this could constitute copyright infringement.

Similarly, AI systems could potentially infringe on intellectual property rights by using code that is protected by trademarks or patents. For example, if an AI system is trained on a training manual that contains code that is protected by a trademark, the AI system could generate code that is identical to or similar to the code in the training manual. If the AI system then uses this code without permission from the trademark owner, this could constitute trademark infringement.

### 19.7.1 Tips for avoiding copyright violations

* Be transparent about what AI tools you use to write your code.
* Obtain permission from the copyright holders of any content that you use to train an AI system. Only use content that has been licensed for use.
* Cite all content that you can.
* Ask the AI tools if the content it helped generate used any content that you can cite.

Did this content use any content from others that I can cite?

## 19.8 Summary

Here is a summary of all the tips we suggested:

* Disclose when you use AI tools to create content.
* Be aware that AI systems are biased and their responses are likely biased. Any content generated by an AI system should be evaluated for potential bias.
* Be aware that AI systems may behave in unexpected ways. Implement new AI solutions slowly to account for the unexpected. Test those systems and try to better understand how they work in different contexts.
* Be aware that humans are still better at generalizing concepts to other contexts.
* Carefully consider if an AI solution is appropriate for your context.
* Credit human authors by citing them and adhering to copyright restrictions.
* Ensure that prompts to commercial tools don’t include private or propriety data or information.
* Cross-check content from AI tools by using multiple AI tools - but check that each tool meets the privacy and security restrictions that you need.
* Don’t assume AI-generated content is real, accurate, consistent, current, or better than that of a human.
* Ask the AI tools to help you understand:
  + Sources for the content that you can cite
  + Any decision processes in how the content was created
  + Potential limitations
  + Potential security or privacy issues
  + Potential downstream consequences of the use of the content
* Always have expert humans review the content and value your own contributions and thoughts.
* Emphasize training and education about AI and recognize that best practices will evolve as the technology evolves.

Overall, we hope that these guidelines and tips will help us all use AI tools more responsibly. We recognize however, that as this is emerging technology and more ethical issues will emerge as we continue to use these tools in new ways. AI tools can even help us to use them more responsibly when we ask the right additional questions, but remember that human review is always necessary. Staying up-to-date on the current ethical considerations will also help us all continue to use AI responsibly.

# 20 Adherence practices

Here we suggest some simple practices that can help you and others at your institution to better adhere to current proposed ethical guidelines.

1. **Start Slow** - Starting slow can allow for time to better understand how AI systems work and any possible unexpected consequences.
2. **Check for Allowed Use**  - AI model responses are often not transparent about using code, text, images and other data types that may violate copyright.
3. **Use Multiple AI Tools** - Using a variety of tools can help reduce the potential for ethical issues that may be specific to one tool, such as bias, misinformation, and security or privacy issues.
4. **Educate Yourself and Others** - To actually comply with ethical standards, it is vital that users be educated about best practices for use. If you help set standards for an institution or group, it strongly advised that you carefully consider how to educate individuals about those standards of use.

Be transparent about what AI tools you use where possible. This help others to better understand how you created any content that was derived by AI, as well as the possible sources that the AI tools might have used when helping you. It may also help with future unknown issues related to the use of these tools.

Keep in mind that some fields, organizations, and societies have guidelines or requirements for using AI, like for example the policy for the use of large language models for the [International Society for Computational Biology](https://www.iscb.org/iscb-policy-statements/iscb-policy-for-acceptable-use-of-large-language-models). Be aware of the requirements/guidelines for your field.

**It is essential to address these ethical concerns and ensure that AI is used in a responsible and transparent manner.** This could be done through ensuring the quality of training for AI systems, promoting transparency about AI-generated content, and implementing safeguards against the creation of harmful or biased content. By doing so, we can harness the potential of AI to improve and transform the way we work while maintaining ethical standards.

**Recognize that the ethical guidelines and standards for your field should be considered when using AI or creating AI use policies.**

## 20.1 Start Slow

Launching large projects using AI before you get a chance to test them could lead to disastrous consequences. Unforeseen challenges have already come up with many uses of AI, so it is wise to start small and do evaluation first before you roll out a system to more users.

This also gives you time to correspond with legal, equity, security, etc. experts about the risks of your AI use.

### 20.1.1 Tips for starting slow

* Consider an early adopters program to evaluate usage.
* Consult with experts about potential unforeseen challenges.
* Continue to assess and evaluate AI systems over time.

## 20.2 Check for Allowed Use

When AI systems are trained on data, they may also learn and incorporate copyrighted information. This means that AI-generated content could potentially infringe on the copyright of the original author. For example, if an AI system is trained on a code written by a human programmer, the AI system could generate code that is identical to or similar to the code from that author. If the AI system then uses this code without permission from the original author, this could constitute copyright infringement.

Similarly, AI systems could potentially infringe on intellectual property rights by using code that is protected by trademarks or patents. For example, if an AI system is trained on a training manual that contains code that is protected by a trademark, the AI system could generate code that is identical to or similar to the code in the training manual. If the AI system then uses this code without permission from the trademark owner, this could constitute trademark infringement.

### 20.2.1 Tips for checking for allowed use

* Be transparent about what AI tools you use to write your code.
* Obtain permission from the copyright holders of any content that you use to train an AI system. Only use content that has been licensed for use.
* Cite all content that you can.
* Ask the AI tools if the content it helped generate used any content that you can cite.

Did this content use any content from others that I can cite?

## 20.3 Use Multiple AI Tools

Only using one AI tool can increase the risk of the ethical issues discussed. For example, it may be easier to determine if a tool incorrect about a response if we see that a variety of tools have different answers to the same prompt. Secondly, as our technology evolves, some tools may perform better than others at specific tasks. It is also necessary to check responses over time with the same tool, to verify that a result is even consistent from the same tool.

### 20.3.1 Tips for using multiple AI tools

* Check that each tool you are using meets the privacy and security restrictions that you need.
* Utilize platforms that make it easier to use multiple AI tools, such as <https://poe.com/>, which as access to many tools, or [Amazon Bedrock](https://aws.amazon.com/about-aws/whats-new/2023/11/evaluate-compare-select-fms-use-case-amazon-bedrock/), which actually has a feature to send the same prompt to multiple tools automatically, including for more advanced usage in the development of models based on modifying existing foundation models.
* Evaluate the results of the same prompt multiple times with the same tool to see how consistent it is overtime.
* Use slightly different prompts to see how the response may change with the same tool.
* Consider if using different types of data maybe helpful for answering the same question.

## 20.4 Educate Yourself and Others

There are many studies indicating that individuals typically want to comply with ethical standards, but it becomes difficult when they do not know how (Giorgini et al. ([2015](#ref-giorgini_researcher_2015))). Furthermore, individuals who receive training are much more likely to adhere to standards (Kowaleski, Sutherland, and Vetter ([2019](#ref-kowaleski_can_2019))).

Properly educating those you wish to comply with standards, can better ensure that compliance actually happens.

It is especially helpful if training materials are developed to be especially relevant to the actually potential uses by the individuals receiving training and if the training includes enough fundamentals so that individuals understand why policies are in place.

**Real World Example**

A lack of proper training at Samsung lead to a leak of proprietary data due to unauthorized use of ChatGPT by employees – see <https://cybernews.com/news/chatgpt-samsung-data-leak> for more details:

“The information employees shared with the chatbot supposedly included the source code of software responsible for measuring semiconductor equipment. A Samsung worker allegedly discovered an error in the code and queried ChatGPT for a solution.

OpenAI explicitly tells users not to share “any sensitive information in your conversations” in the company’s frequently asked questions (FAQ) section. Information that users directly provide to the chatbot is used to train the AI behind the bot.

Samsung supposedly discovered three attempts during which confidential data was revealed. Workers revealed restricted equipment data to the chatbot on two separate occasions and once sent the chatbot an excerpt from a corporate meeting. Privacy concerns over ChatGPT’s security have been ramping up since OpenAI revealed that a flaw in its bot exposed parts of conversations users had with it, as well as their payment details in some cases. As a result, the Italian Data Protection Authority has banned ChatGPT, while German lawmakers have said they could follow in Italy’s footsteps.”

### 20.4.1 Tips to educate yourself and others

* Emphasize the importance of training and education
* Recognize that general AI literacy to better understand how AI works, can help individuals use AI more responsibly.
* Seek existing education content made by experts that can possibly be modified for your use case
* Consider how often people will need to be reminded about best practices. Should training be required regularly? Should individuals receive reminders about best practices especially in contexts in which they might use AI tools.
* Make your best practices easily findable and help point people to the right individuals to ask for guidance.
* Recognize that best practices for AI will likely change frequently in the near future as the technology evolves, education content should be updated accordingly.

## 20.5 Summary

Here is a summary of all the tips we suggested:

* Disclose when you use AI tools to create content.
* Be aware that AI systems may behave in unexpected ways. Implement new AI solutions slowly to account for the unexpected. Test those systems and try to better understand how they work in different contexts.
* Adhere to copyright restrictions for use of data and content created by AI systems.
* Cross-check content from AI tools by using multiple AI tools and checking for consistent results over time. Check that each tool meets the privacy and security restrictions that you need.
* Emphasize training and education about AI and recognize that best practices will evolve as the technology evolves.

Overall, we hope that these suggestions will help us all use AI tools more responsibly. We recognize however, that as this is emerging technology and more ethical issues will emerge as we continue to use these tools in new ways. AI tools can even help us to use them more responsibly when we ask the right additional questions, but remember that human review is always necessary. Staying up-to-date on the current ethical considerations will also help us all continue to use AI responsibly.

# 21 Consent and AI

ChatGPT example for Samsung or use that earlier?

# 22 IDARE and AI

IDARE stands for Inclusion, Diversity, Anti-Racism, and Equity. It is an acronym used by some institutions (such as the [Johns Hopkins Bloomberg School of Public Health](https://publichealth.jhu.edu/offices-and-services/office-of-inclusion-diversity-anti-racism-and-equity-idare), [the University of California, Davis](https://health.ucdavis.edu/diversity-inclusion/committees/departmental-committees-index.html), and the [University of Pennsylvania Perelman School of Medicine](https://www.med.upenn.edu/neurology/idare/)) to remind people about practices to improve [social justice](https://en.wikipedia.org/wiki/Social_justice). As we strive to use AI responsibly, keeping the major principles of IDARE in mind will be helpful to better ensure that individuals of all backgrounds and life experiences more equally benefit from advances in technological and that technology is not used to perpetuate harm.

## 22.1 AI is biased

Humans are biased, therefore data from text written by humans is often also biased, which mean AI systems built on human text are trained to be biased, even those created with the best intentions (Pethig and Kroenung ([2023](#ref-pethig_biased_2023))).

To better understand your own personal bias, consider taking a test at <https://implicit.harvard.edu/>.

It is nearly impossible to create a training dataset that is free from all possible bias and include all possible example data, so by necessity the data used to train AI systems are generally biased in some way and lack data about people across the full spectrum of backgrounds and life experiences. This can lead to AI-created products that cause discrimination, abuse, or neglect for certain groups of people, such as those with certain ethnic or cultural backgrounds, genders, ages, sexuality, capabilities, religions or other group affiliations. Our goal is to create and use AI systems that are as inclusive and unbiased as possible while also keeping in mind that the system is not perfect.

To learn more about how AI algorithms become biased, see <https://www.criticalracedigitalstudies.com/peoplesguide>.

## 22.2 Examples of AI Bias

There are many examples in which biased AI systems were used in a context with negative consequences.

### 22.2.1 Amazon’s resume system was biased against women

Amazon used an AI system was to help filter candidates for jobs. They started using the system in 2014. In 2015, it was discovered that the system penalized resumes that included words like “women’s”, and also for graduates of two all-women’s colleges (Dastin ([2018](#ref-dastin_insight_2018))).

How did this happen? The model was trained on resume’s of existing Amazon employees and most of their employees were male. Thus the training data for this system was not gender inclusive, which lead to bias in the model.

### 22.2.2 X-ray studies show AI surprises

Algorithms used to evaluate medical images seem to be predicting the self-reported race of the individuals in the images from the images alone (Gichoya et al. ([2022](#ref-gichoya_ai_2022))). This is despite the fact that the radiologists examining those same images were not able to identify what aspect about the images helped the AI systems identify the race of the individuals. Why is this a problem? That information from models that evaluate medical images are being used to help suggest care. It is recognized that health disparities exist in the treatment of different racial groups. Therefore bias related to these disparities may be perpetuated by algorithms even when the AI system is trained in a manner that is “blind” to the self-reported race of the individuals.

This example shows that AI systems can possibly amplify existing biases even when humans are unaware of the AI systems using those biases to make decisions.

This is especially a problem, as some populations are under-diagnosed and therefore denied care or they receive poorer care because an AI system does not work as well for their population (Ricci Lara, Echeveste, and Ferrante ([2022](#ref-ricci_lara_addressing_2022))). As an example, a study evaluating diagnosis of various diseases from chest X-ray images, found that certain groups of patients, such as females, those under 20, those who self report as Black or Hispanic, were more likely to be falsely flagged by AI system as healthy when they in fact had an issue (Seyyed-Kalantari et al. ([2021](#ref-seyyed-kalantari_underdiagnosis_2021))). Another example shows that processing of cardiac images from specific patient populations is much poorer using models where the training set was not diverse enough (Puyol-Anton et al. ([2021](#ref-puyol-anton_fairness_2021))).

However, there is promise for good AI systems to mitigate bias. For example, a team studying pain levels in osteoarthritis (a disease where under-served populations often have higher than expected levels of pain) found that using predictions of pain based on AI system examining images were much more accurate than predictions from radiologists examining those same images (Pierson et al. ([2021](#ref-pierson_algorithmic_2021))).

A magazine article describing this work stated:

In this case, researchers were training the models based on physician reports of pain, and since doctors are less likely to believe marginalized people when they report pain, this algorithm replicated this bias. When a team of computer scientists at the University of California, Berkeley, tweaked the algorithm to factor in patient pain reports rather than a physician’s, however, they eliminated that racial bias, paving the way for more equitable treatment of osteoarthritis.” (Arnold ([2022](#ref-Arnold_23)))

## 22.3 Mitigation

When working with AI systems, users should actively identify any potential biases used in the training data for a particular AI system. In particular, the user should look for harmful data, such as discriminatory and false associations included in the training dataset, as well as verify whether the training data is adequately inclusive for your needs. A lack of data about certain ethnic or gender groups or disabled individuals could result in a product that does not adequately consider these groups, ignores them all together, or makes false associations. Where possible, users of commercial AI tools should ask prompts in a manner that includes concern for equity and inclusion, they should use tools that are transparent about what training data was used and limitations of this data, and they should always question the responses from the tool for possible bias.

Why did you assume that the individual was male?

Those developing or augmenting models should also evaluate the training data and the model for biases and false associations as it is being developed instead of waiting to test the product after creation is finished. This includes verifying that the product works properly for potential use cases from a variety of ethnic, gender, ability, socioeconomic, language, and educational backgrounds. When possible, the user should also augment the training dataset with data from groups that are missing or underrepresented in the original training dataset.

## 22.4 Be extremely careful using AI for decisions

There is a common misconception that AI tools might make better decisions for humans because they are believed to not be biased like humans (Pethig and Kroenung ([2023](#ref-pethig_biased_2023))). However since they are built by humans and trained on human data, they are also biased. It is possible that AI systems specifically trained to avoid bias, to be inclusive, to be anti-racist, and for specific contexts may be helpful to enable a more neutral party, but that is generally not currently possible.

AI should not be used to make or help make employment decisions about applicants or employees at this time. This includes recruitment, hiring, retention, promotions, transfers, performance monitoring, discipline, demotion, terminations, or other similar decisions.

## 22.5 More inclusive teams means better models

It is vital that teams hired for the development, auditing or testing of AI tools be as inclusive as possible and should follow the current best IDARE practices for standards for hiring standards. This will help to ensure that different perspectives and concerns are considered.

## 22.6 Access

Resources:

<https://arxiv.org/abs/2311.14096>

<https://www3.weforum.org/docs/WEF_A_Blueprint_for_Equity_and_Inclusion_in_Artificial_Intelligence_2022.pdf>

<https://magazine.jhsph.edu/2022/how-biased-data-and-algorithms-can-harm-health>

<https://research.csiro.au/ss/science/projects/responsible-ai-pattern-catalogue/rai-standard/>

<https://www.tandfonline.com/doi/full/10.1080/08839514.2023.2176618>

# 23 Introduction to Establishing AI Infrastructure

## 23.1 Motivation

## 23.2 Target Audience

The course is intended for …

## 23.3 Curriculum

The course covers…

Navigate the Evolving Legal Landscape of AI with Confidence

# 24 Introduction to AI Policy

In today’s rapidly evolving technological landscape, staying ahead of regulations can be challenging, especially when dealing with cutting-edge technologies like artificial intelligence (AI). This course equips decision-makers and leaders with the knowledge they need to navigate the legal complexities surrounding AI.

## 24.1 Motivation

By understanding the key legal considerations and emerging regulations, you can ensure your organization leverages the power of AI while staying compliant and mitigating potential risks. This course empowers you to make informed decisions and confidently utilize AI to achieve your organizational goals.

## 24.2 Target Audience

This course is targeted toward industry and non-profit leaders and decision makers.

## 24.3 Curriculum

In this course, we’ll learn about the state of existing AI laws, other laws and regulations that can apply to AI systems and products, how to create a strong AI policy for your organization, and how to develop compliance training for your workforce.

# 25 Legal Regulations Around AI Use

Because generative AI is a fairly new technology for most fields, the regulations and laws surrounding its use are in flux and changing rapidly. A provisional deal on one of the first major set of AI regulations was announced by the EU government on Friday, December 8, 2023 (the [AI Act](https://artificialintelligenceact.eu/documents/)). These rules will apply to AI regulation and use within the 27-member EU bloc, as well as to foreign companies that operate within the EU, making it likely EU AI Act will guide regulations around the globe.

Countries outside of the EU are drafting their own laws and standards surrounding AI use, so you will need to do some research on what it and is not allowed in your local area.

<>

## 25.1 The EU AI Act

According to EU policymakers involved in writing the AI Act, the goal of the Act is to regulate AI in order to limit its capacity to cause harm. The political agreement covers the use of AI in biometric surveillance (such as facial recognition), as well as guidance on regulations for LLMs. The EU AI Act divides AI-based products into levels based on how much risk each product might pose to things like data privacy and protection. Higher-risk products with a greater capacity to cause harm face more stringent rules and regulations.

Final details are still being worked out, but we do know several important aspects of this Act.

1. All content generated by AI must be clearly identified.
2. Foundational models like GPT as well as general purpose AI systems (GPAI) must create technical documentation and detailed summaries about the training data before they can be released on the market.
3. **High-risk AI systems** must undergo mandatory rights impact and mitigation assessments. Developers will also have to conduct model evaluations, assess and track possible cybersecurity risks, and report serious incidents and breaches to the European Commission.
4. **Open-source software** is excluded from regulations, with some exceptions for software that is considered a high-risk system or is a prohibited application.
5. AI software for manipulative strategies like deepfakes and automated disinformation campaigns, systems exploiting vulnerabilities, and indiscriminate scraping of facial images from the internet or security footage to create facial recognition databases are banned. Additional prohibited applications may be added later.
6. There are exceptions to the facial scraping ban that allow law enforcement and intelligence agencies to use AI applications for facial recognition purposes.

The AI Act also lays out financial penalties for companies that violate these regulations, which can be as high as 7% of a company’s global revenue.

More information about the EU’s AI Act can be found in these sources. <<>> <https://www.washingtonpost.com/technology/2023/12/08/ai-act-regulation-eu/>

<https://www.reuters.com/technology/stalled-eu-ai-act-talks-set-resume-2023-12-08/>

<https://www.euractiv.com/section/artificial-intelligence/news/ai-act-eu-policymakers-nail-down-rules-on-ai-models-butt-heads-on-law-enforcement/>

<https://www.ciodive.com/news/EU-AI-Act-penalties-guardrails-foundational-models/702192/#>:~:text=The%20EU’s%20proposed%20regulations%20prohibit,free%20will%20and%20using%20the

# 26 VIDEO Legal Regulations Around AI Use

# 27 Existing Laws That Apply to AI

While countries and jurisdictions are developing ans passing laws that specifically deal with AI, there are also existing laws around data that should be considered whenever working with AI. These broadly include regulations about intellectual property, data privacy and protection, and liability.

## 27.1 Intellectual Property Laws

There are multiple concerns around generative AI and intellectual property rights. These include potential copyright violations in the training data, whether generative AI methods can be proprietary, and whether material created using generative AI can be copyrighted. In order for generative AI models to work, they must be trained on vast amounts of data. This might include images, in the case of image generators like DALL-E, Stable Diffusion, and Midjourney. It might also include human writing and speech, in the case of LLMs like ChatGPT and Bard. Information about the training data sets for these tools is limited, but they likely include text and images scraped from the internet. There is concern that the text and images gathered for training data included copyrighted and trademarked books, articles, photographs, and artwork. In fact, the CEO of Midjourney confirmed that copyrighted images were included in the Midjourney training data without the consent of the artists, and there was no way for artists to opt out of having their work included. A group of authors have also recently sued OpenAI, the company behind ChatGPT, for copyright infringement because their published works were included in the GPT training data. Additionally, some artists have sued AI companies like the company behind Stable Diffusion because the generative AI tools are creating images that are too similar to their existing, protected artwork.

There is also ongoing debate as to whether AI-generated images and text can be copyrighted. While many current copyright laws do not protect works created by machines, how these laws might apply to work that is a collaboration between humans and machine (such as art that includes some AI-generated content) is an area of active discussion.

## 27.2 Data Privacy and Protection

### 27.2.1 Information Security

### 27.2.2 Personal Information

Data privacy is especially important to consider when working in fields like healthcare, biomedical research, and education, where personally identifiable data and personal health information is under special protections.

HIPAA?

GPAI?

### 27.2.3 Biometric Privacy Laws

Biometric data involves human characteristics gathered from physical or behavioral traits that can be used to identify a single person. This might include things like fingerprints, palm prints, iris scans, facial scans, and voice recognition. DNA can also be considered biometric data when used for forensics.

## 27.3 Liability Laws

As AI systems become more and more common in everyday life, it is inevitable that some of these systems will fail at some point. Who is liable when AI fails, especially when it fails in a catastrophic manner?

The issue of whose fault it is when an AI system fails (and thus who is responsible for the damage) depends greatly on *how* and *why* it failed. Blame might lie with the user (if the AI was not being used according to instructions, or if limitations were known but ignored), the software developer (if the AI product was distributed before being tested thoroughly or before the algorithm was properly tuned), or the designer or manufacturer (if the AI design or production was inherently flawed).

# 28 VIDEO Existing Laws That Apply to AI

# 29 Title!

Everyone needs an AI policy now

# 30 VIDEO Title!

# 31 Title!

“Making it easy for people to succeed in compliance”

# 32 VIDEO Title!

# About the Authors

These credits are based on our [course contributors table guidelines](https://www.ottrproject.org/more_features.html#giving-credits-to-contributors).

| Credits | Names |
| --- | --- |
| **Pedagogy** |  |
| Lead Content Instructor(s) | [FirstName LastName](link%20to%20personal%20website) |
| Lecturer(s) (include chapter name/link in parentheses if only for specific chapters) - make new line if more than one chapter involved | Delivered the course in some way - video or audio |
| Content Author(s) (include chapter name/link in parentheses if only for specific chapters) - make new line if more than one chapter involved | If any other authors besides lead instructor |
| Content Contributor(s) (include section name/link in parentheses) - make new line if more than one section involved | Wrote less than a chapter |
| Content Editor(s)/Reviewer(s) | Checked your content |
| Content Director(s) | Helped guide the content direction |
| Content Consultants (include chapter name/link in parentheses or word “General”) - make new line if more than one chapter involved | Gave high level advice on content |
| Acknowledgments | Gave small assistance to content but not to the level of consulting |
| **Production** |  |
| Content Publisher(s) | Helped with publishing platform |
| Content Publishing Reviewer(s) | Reviewed overall content and aesthetics on publishing platform |
| **Technical** |  |
| Course Publishing Engineer(s) | Helped with the code for the technical aspects related to the specific course generation |
| Template Publishing Engineers | [Candace Savonen](https://www.cansavvy.com/), [Carrie Wright](https://carriewright11.github.io/), [Ava Hoffman](https://www.avahoffman.com/) |
| Publishing Maintenance Engineer | [Candace Savonen](https://www.cansavvy.com/) |
| Technical Publishing Stylists | [Carrie Wright](https://carriewright11.github.io/), [Ava Hoffman](https://www.avahoffman.com/), [Candace Savonen](https://www.cansavvy.com/) |
| Package Developers ([ottrpal](https://github.com/jhudsl/ottrpal)) [Candace Savonen](https://www.cansavvy.com/), [John Muschelli](https://johnmuschelli.com/), [Carrie Wright](https://carriewright11.github.io/) |  |
| **Art and Design** |  |
| Illustrator(s) | Created graphics for the course |
| Figure Artist(s) | Created figures/plots for course |
| Videographer(s) | Filmed videos |
| Videography Editor(s) | Edited film |
| Audiographer(s) | Recorded audio |
| Audiography Editor(s) | Edited audio recordings |
| **Funding** |  |
| Funder(s) | Institution/individual who funded course including grant number |
| Funding Staff | Staff members who help with funding |

## ─ Session info ───────────────────────────────────────────────────────────────  
## setting value   
## version R version 4.0.2 (2020-06-22)  
## os Ubuntu 20.04.5 LTS   
## system x86\_64, linux-gnu   
## ui X11   
## language (EN)   
## collate en\_US.UTF-8   
## ctype en\_US.UTF-8   
## tz Etc/UTC   
## date 2023-12-14   
##   
## ─ Packages ───────────────────────────────────────────────────────────────────  
## package \* version date lib source   
## assertthat 0.2.1 2019-03-21 [1] RSPM (R 4.0.5)   
## bookdown 0.24 2023-03-28 [1] Github (rstudio/bookdown@88bc4ea)   
## cachem 1.0.7 2023-02-24 [1] CRAN (R 4.0.2)   
## callr 3.5.0 2020-10-08 [1] RSPM (R 4.0.2)   
## cli 3.6.1 2023-03-23 [1] CRAN (R 4.0.2)   
## crayon 1.3.4 2017-09-16 [1] RSPM (R 4.0.0)   
## desc 1.2.0 2018-05-01 [1] RSPM (R 4.0.3)   
## devtools 2.3.2 2020-09-18 [1] RSPM (R 4.0.3)   
## digest 0.6.25 2020-02-23 [1] RSPM (R 4.0.0)   
## ellipsis 0.3.1 2020-05-15 [1] RSPM (R 4.0.3)   
## evaluate 0.20 2023-01-17 [1] CRAN (R 4.0.2)   
## fastmap 1.1.1 2023-02-24 [1] CRAN (R 4.0.2)   
## fs 1.5.0 2020-07-31 [1] RSPM (R 4.0.3)   
## glue 1.4.2 2020-08-27 [1] RSPM (R 4.0.5)   
## htmltools 0.5.5 2023-03-23 [1] CRAN (R 4.0.2)   
## knitr 1.33 2023-03-28 [1] Github (yihui/knitr@a1052d1)   
## magrittr 2.0.3 2022-03-30 [1] CRAN (R 4.0.2)   
## memoise 2.0.1 2021-11-26 [1] CRAN (R 4.0.2)   
## pkgbuild 1.1.0 2020-07-13 [1] RSPM (R 4.0.2)   
## pkgload 1.1.0 2020-05-29 [1] RSPM (R 4.0.3)   
## prettyunits 1.1.1 2020-01-24 [1] RSPM (R 4.0.3)   
## processx 3.4.4 2020-09-03 [1] RSPM (R 4.0.2)   
## ps 1.4.0 2020-10-07 [1] RSPM (R 4.0.2)   
## R6 2.4.1 2019-11-12 [1] RSPM (R 4.0.0)   
## remotes 2.2.0 2020-07-21 [1] RSPM (R 4.0.3)   
## rlang 1.1.0 2023-03-14 [1] CRAN (R 4.0.2)   
## rmarkdown 2.10 2023-03-28 [1] Github (rstudio/rmarkdown@02d3c25)  
## rprojroot 2.0.3 2022-04-02 [1] CRAN (R 4.0.2)   
## sessioninfo 1.1.1 2018-11-05 [1] RSPM (R 4.0.3)   
## stringi 1.5.3 2020-09-09 [1] RSPM (R 4.0.3)   
## stringr 1.4.0 2019-02-10 [1] RSPM (R 4.0.3)   
## testthat 3.0.1 2023-03-28 [1] Github (R-lib/testthat@e99155a)   
## usethis 1.6.3 2020-09-17 [1] RSPM (R 4.0.2)   
## withr 2.3.0 2020-09-22 [1] RSPM (R 4.0.2)   
## xfun 0.26 2023-03-28 [1] Github (yihui/xfun@74c2a66)   
## yaml 2.2.1 2020-02-01 [1] RSPM (R 4.0.3)   
##   
## [1] /usr/local/lib/R/site-library  
## [2] /usr/local/lib/R/library

# 33 References

Arnold, Carrie. 2022. “How Biased Data and Algorithms Can Harm Health Hopkins Bloomberg Public Health Magazine.” <https://magazine.jhsph.edu/2022/how-biased-data-and-algorithms-can-harm-health>.

Belenguer, Lorenzo. 2022. “AI Bias: Exploring Discriminatory Algorithmic Decision-Making Models and the Application of Possible Machine-Centric Solutions Adapted from the Pharmaceutical Industry.” *Ai and Ethics* 2 (4): 771–87. <https://doi.org/10.1007/s43681-022-00138-8>.

Cote, Catherine. 2022. “7 DATA COLLECTION METHODS IN BUSINESS ANALYTICS.” <https://online.hbs.edu/blog/post/data-collection-methods>.

Dastin, Jeffrey. 2018. “Insight - Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women.” *Reuters*, October. <https://www.reuters.com/article/idUSKCN1MK0AG/>.

Gichoya, Judy Wawira, Imon Banerjee, Ananth Reddy Bhimireddy, John L. Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, et al. 2022. “AI Recognition of Patient Race in Medical Imaging: A Modelling Study.” *The Lancet Digital Health* 4 (6): e406–14. <https://doi.org/10.1016/S2589-7500(22)00063-2>.

Giorgini, Vincent, Jensen T. Mecca, Carter Gibson, Kelsey Medeiros, Michael D. Mumford, Shane Connelly, and Lynn D. Devenport. 2015. “Researcher Perceptions of Ethical Guidelines and Codes of Conduct.” *Accountability in Research* 22 (3): 123–38. <https://doi.org/10.1080/08989621.2014.955607>.

Hamzelou, Jessica. n.d. “Artificial Intelligence Is Infiltrating Health Care. We Shouldn’t Let It Make All the Decisions.” *MIT Technology Review*. Accessed May 8, 2023. <https://www.technologyreview.com/2023/04/21/1071921/ai-is-infiltrating-health-care-we-shouldnt-let-it-make-decisions/>.

Kowaleski, Zachary T., Andrew Sutherland, and Felix Vetter. 2019. “Can Ethics Be Taught? Evidence from Securities Exams and Investment Adviser Misconduct.” {SSRN} {Scholarly} {Paper}. Rochester, NY. <https://doi.org/10.2139/ssrn.3457588>.

Mattson, Mark P. 2014. “Superior Pattern Processing Is the Essence of the Evolved Human Brain.” *Frontiers in Neuroscience* 8. <https://doi.org/10.3389/fnins.2014.00265>.

Nikulski, Julia. 2021. “Toxicity in AI Text Generation.” *Medium*. <https://towardsdatascience.com/toxicity-in-ai-text-generation-9e9d9646e68f>.

Paul, Amy K, and Merrick Schaefer. 2020. “Safeguards for the Use of Artificial Intelligence and Machine Learning in Global Health.” *Bulletin of the World Health Organization* 98 (4): 282–84. <https://doi.org/10.2471/BLT.19.237099>.

Pearce, Guy. 2021. “Beware the Privacy Violations in Artificial Intelligence Applications.” *ISACA*. <https://www.isaca.org/resources/news-and-trends/isaca-now-blog/2021/beware-the-privacy-violations-in-artificial-intelligence-applications>.

Pethig, Florian, and Julia Kroenung. 2023. “Biased Humans, (Un)Biased Algorithms?” *Journal of Business Ethics* 183 (3): 637–52. <https://doi.org/10.1007/s10551-022-05071-8>.

Pierson, Emma, David M. Cutler, Jure Leskovec, Sendhil Mullainathan, and Ziad Obermeyer. 2021. “An Algorithmic Approach to Reducing Unexplained Pain Disparities in Underserved Populations.” *Nature Medicine* 27 (1): 136–40. <https://doi.org/10.1038/s41591-020-01192-7>.

Puyol-Anton, Esther, Bram Ruijsink, Stefan K. Piechnik, Stefan Neubauer, Steffen E. Petersen, Reza Razavi, and Andrew P. King. 2021. “Fairness in Cardiac MR Image Analysis: An Investigation of Bias Due to Data Imbalance in Deep Learning Based Segmentation.” arXiv. <https://doi.org/10.48550/arXiv.2106.12387>.

Ricci Lara, María Agustina, Rodrigo Echeveste, and Enzo Ferrante. 2022. “Addressing Fairness in Artificial Intelligence for Medical Imaging.” *Nature Communications* 13 (1): 4581. <https://doi.org/10.1038/s41467-022-32186-3>.

Seyyed-Kalantari, Laleh, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen, and Marzyeh Ghassemi. 2021. “Underdiagnosis Bias of Artificial Intelligence Algorithms Applied to Chest Radiographs in Under-Served Patient Populations.” *Nature Medicine* 27 (12): 2176–82. <https://doi.org/10.1038/s41591-021-01595-0>.

Walsh, Dylan. 2023. “The Legal Issues Presented by Generative AI.” <https://mitsloan.mit.edu/ideas-made-to-matter/legal-issues-presented-generative-ai>.