



# Ethical AI microcredential

BOOKLET

**CU1 | What is Algorithmic Bias?**

# How to use this Flipbook?

This document is interactive. Throughout the document, you will find links to additional information.



Button that takes you to the beginning of the document. This icon appears on the top right corner of the pages.



Whenever you see this arrow, it means that you have an **interactive color text** to click on, that has an external link associated to it.

**DISCLAIMER:** Please note that we cannot guarantee the continued availability of external content, such as videos, as they may be subject to change or removal by its authors or host platforms.

# Index

Click on the menu

**01. Introduction**

**02. Course contents and expected outputs**

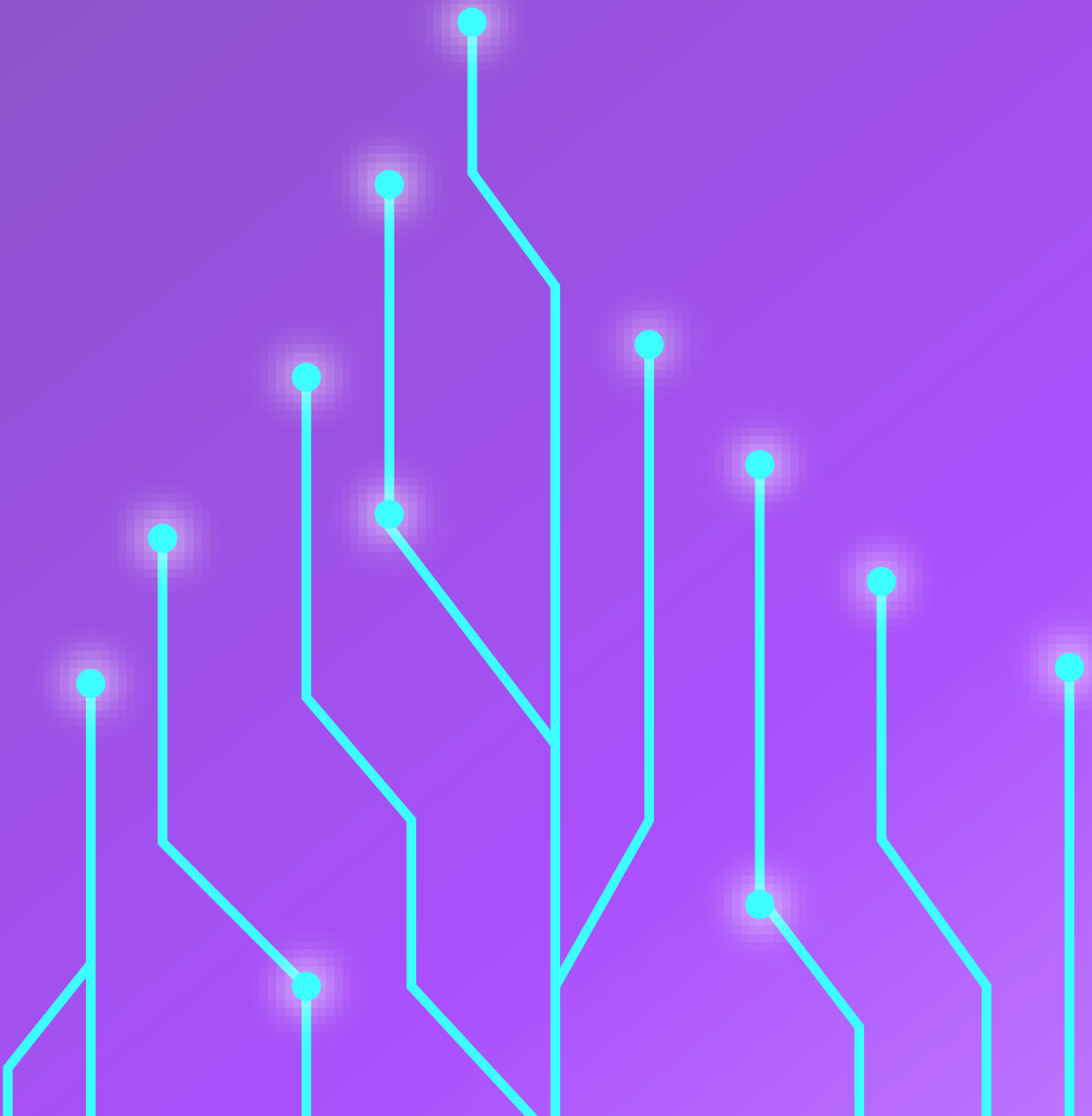
**03. What is Algorithmic Bias?**

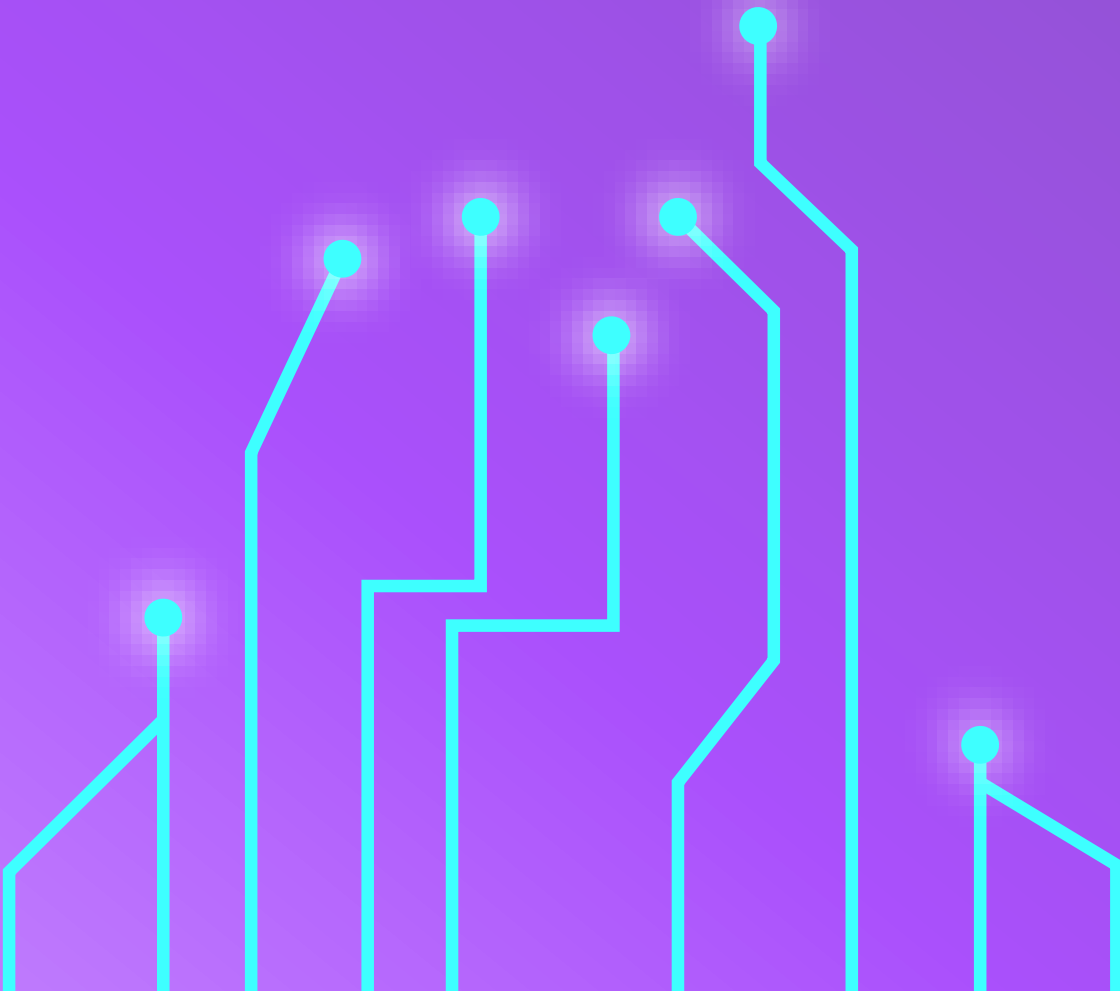
**04. Defining Algorithmic Bias**

**05. Understanding Bias in AI Systems**

# 01. Introduction

CU1 | What is Algorithmic Bias?





## 01. Introduction

In the rapidly evolving landscape of artificial intelligence (AI), ensuring ethical development and use of AI technologies is paramount. This booklet serves as your comprehensive guide to the Ethical AI microcredential, focusing on six competence units designed to equip you with the knowledge and skills necessary to navigate the ethical complexities of AI.

As you embark on this journey, you will explore six distinct competence units, each addressing crucial aspects of ethical AI development and deployment. From understanding algorithmic bias to promoting transparency and upholding human rights, these competence units are designed to empower you with the tools needed to navigate the ethical challenges inherent in AI technologies.

Throughout this booklet, you will delve into the following Competence Units (CU from now on):

- CU1 - What is Algorithmic Bias?
- CU2 - Non-maleficence
- CU3 - Accountability
- CU4 - Transparency
- CU5 - Human Rights and Fairness
- CU6 - AI Ethics, a Practical Approach



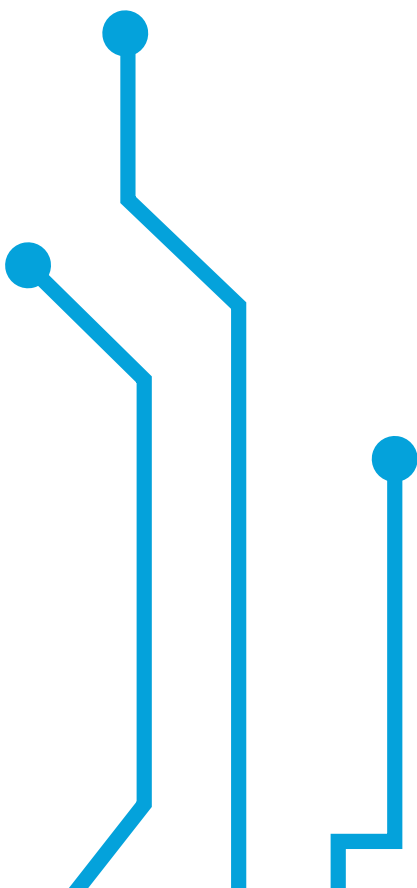
Each unit will provide you with a deeper understanding of key ethical principles and practices in AI, along with practical insights and real-world examples to reinforce your learning.

Whether you are an adult student, a professional, or an AI enthusiast, this booklet offers a valuable resource to expand your knowledge and expertise in Ethical AI. We invite you to embark on this journey with us as we explore the ethical dimensions of AI and work towards creating a more responsible and equitable future.

Thank you for choosing this booklet as your guide to ethical AI development and practice.

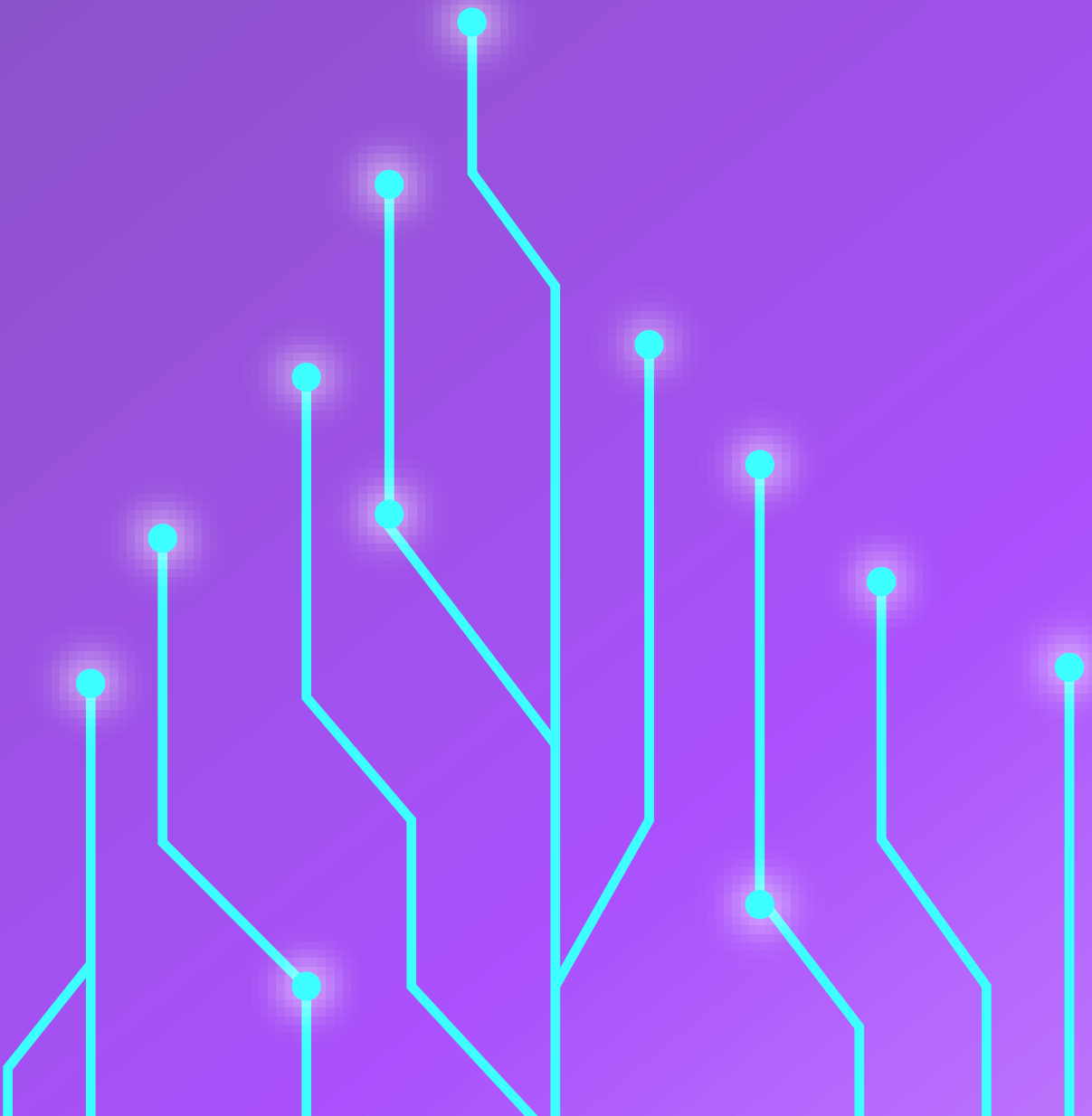
Let's begin this transformative journey together!

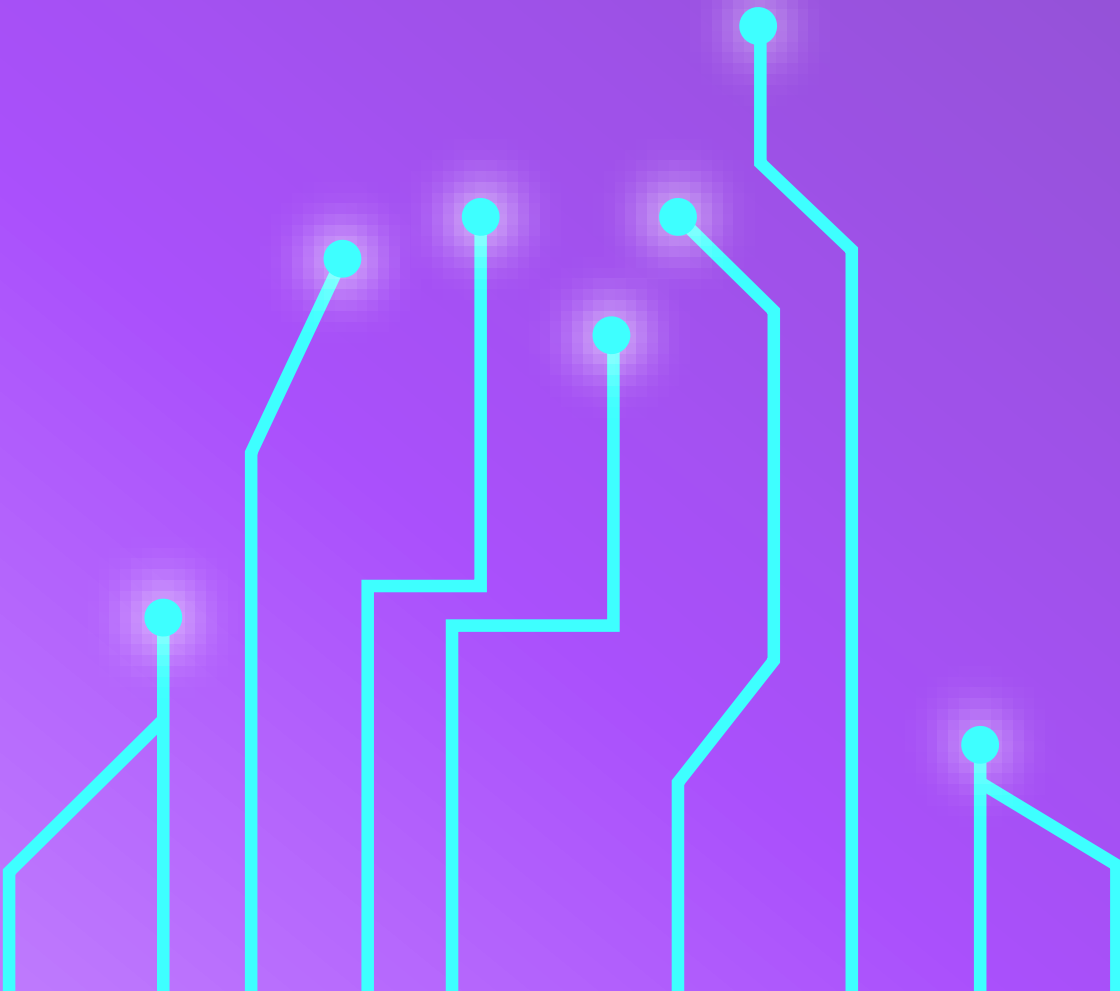
*The CHARLIE project team*



# 02. Course contents and expected outputs

CU1 | What is Algorithmic Bias?





## 02. Course contents and expected outputs

The "Ethical AI Microcredential" at EQF4 level is designed to achieve the following **outcomes**:

1. Establish a **foundational understanding of algorithmic bias**, exploring its origins and implications for individuals and society.
  - Delve into the definition, sources, and manifestations of algorithmic bias.
  - Analyse the societal and individual ramifications of biased algorithms.
2. Cultivate awareness and application of the **ethical principle of non-maleficence in AI** development.
  - Assess the risks and harms associated with biased algorithms.
  - Develop strategies to mitigate harm and promote ethical AI development.
3. Appreciate the significance of **accountability in AI systems**, examining pertinent legal and ethical frameworks.
  - Investigate the roles of various stakeholders in AI accountability.
  - Learn best practices for fostering accountability in AI development.



4. Gain insight into the concept of **transparency in AI systems** and its pivotal role in algorithmic decision-making.
  - Explore methodologies and tools to enhance transparency in AI.
  - Understand the challenges and constraints in rendering complex algorithms more comprehensible.
5. Explore the **intersection of AI, human rights, and fairness**, and their implications for ethical AI development.
  - Assess the impact of biased algorithms on human rights, including non-discrimination, privacy, and freedom of expression.
  - Develop strategies to ensure fairness and equity in AI development and deployment.
6. Apply ethical principles in AI development and deployment through **practical approaches and real-world scenarios**.
  - Examine various ethical frameworks and guidelines and their application to AI systems.
  - Understand the importance of stakeholder engagement, interdisciplinary collaboration, and ethical AI development processes.



Upon completion of this course, participants will possess a holistic understanding of algorithmic biases, their sector-specific impacts, and the tools and strategies to address them. This knowledge equips professionals and academics/students in algorithm-driven fields to contribute to more equitable and fair outcomes in a data-driven world.

The microcredential course is structured around **6 Competency Units (CU's)**, each designed to equip participants with the knowledge and skills necessary to navigate the challenges and opportunities in the realm of algorithmic bias.

**CU1 – What is Algorithmic Bias?** In this unit, students will explore the concept of algorithmic bias and its various manifestations. It covers its definition, causes, and societal implications. Students will analyse bias origins, sources within algorithms, and potential impacts on individuals and society.

**CU2 – Non-maleficence:** This unit delves into the ethical principle of non-maleficence, prioritising the avoidance of harm in AI development and deployment. Participants will explore the inherent risks and harms linked to biased algorithms, while also uncovering strategies to mitigate these risks and foster ethical AI practices.



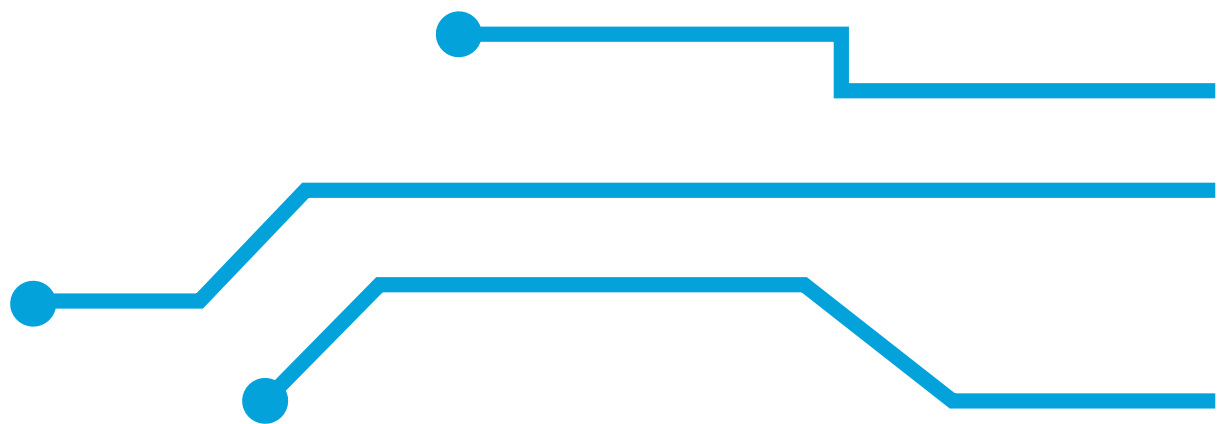
**CU3 – Accountability:** In this unit, students delve into the critical realm of accountability within AI development and utilisation. Expounding on the necessity of delineating clear lines of responsibility, participants explore legal and ethical frameworks governing accountability. Additionally, the curriculum scrutinises the roles of diverse stakeholders and delves into best practices ensuring accountability throughout AI development endeavours.

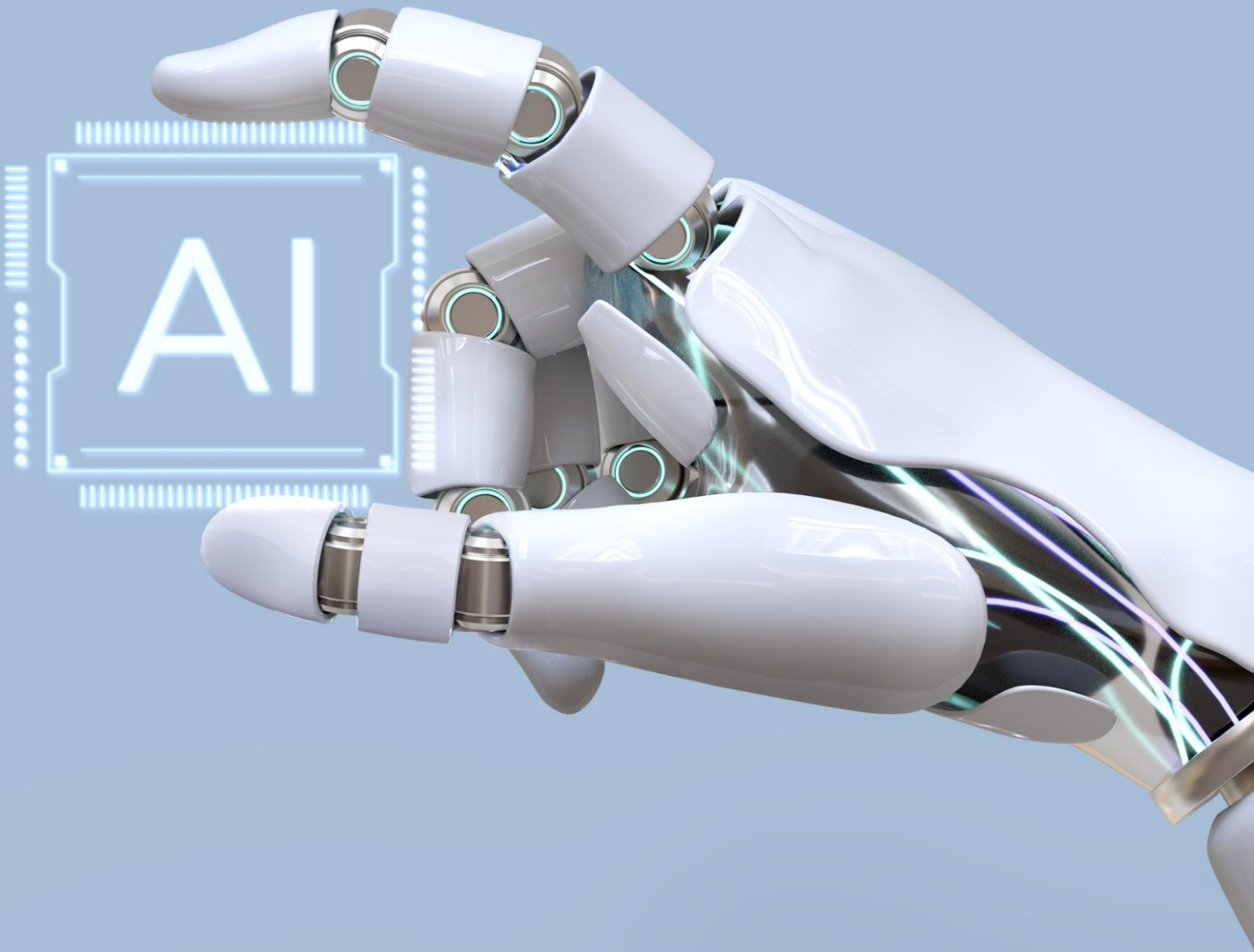
**CU4 – Transparency:** This unit illuminates the significance of transparency within AI systems, emphasising the values of openness, communication, and explainability in algorithmic decision-making. Participants will engage with techniques and resources aimed at augmenting transparency in AI, while also grappling with the inherent challenges and constraints in rendering intricate algorithms comprehensible.

**CU5 – Human rights and fairness:** In the Human rights and fairness unit, students will explore the intersection of AI, human rights, and fairness. They will examine how biased algorithms can impact human rights, including the right to non-discrimination, privacy, and freedom of expression. Students will also learn about strategies for ensuring fairness and equity in AI development and deployment.

**CU6 - AI Ethics, a practical approach:** This unit emphasises the pragmatic application of ethical principles throughout AI development and deployment. Participants delve into diverse ethical frameworks and guidelines, gaining insight into their real-world application in AI scenarios. Additionally, the unit underscores the significance of stakeholder engagement, interdisciplinary collaboration, and the integration of ethical AI development processes for fostering responsible AI innovation.

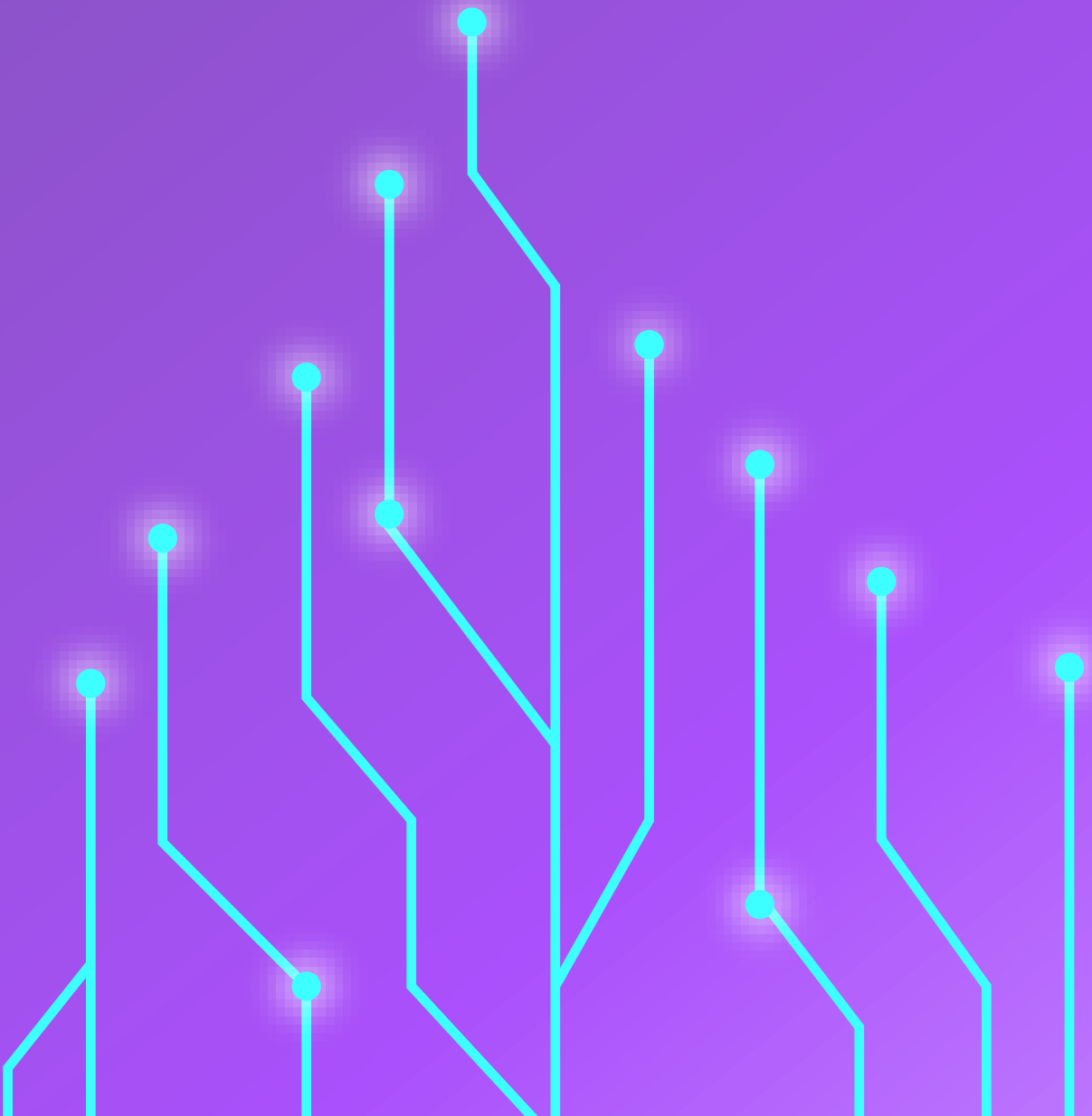
The subsequent section elaborates on the content of each competence unit.

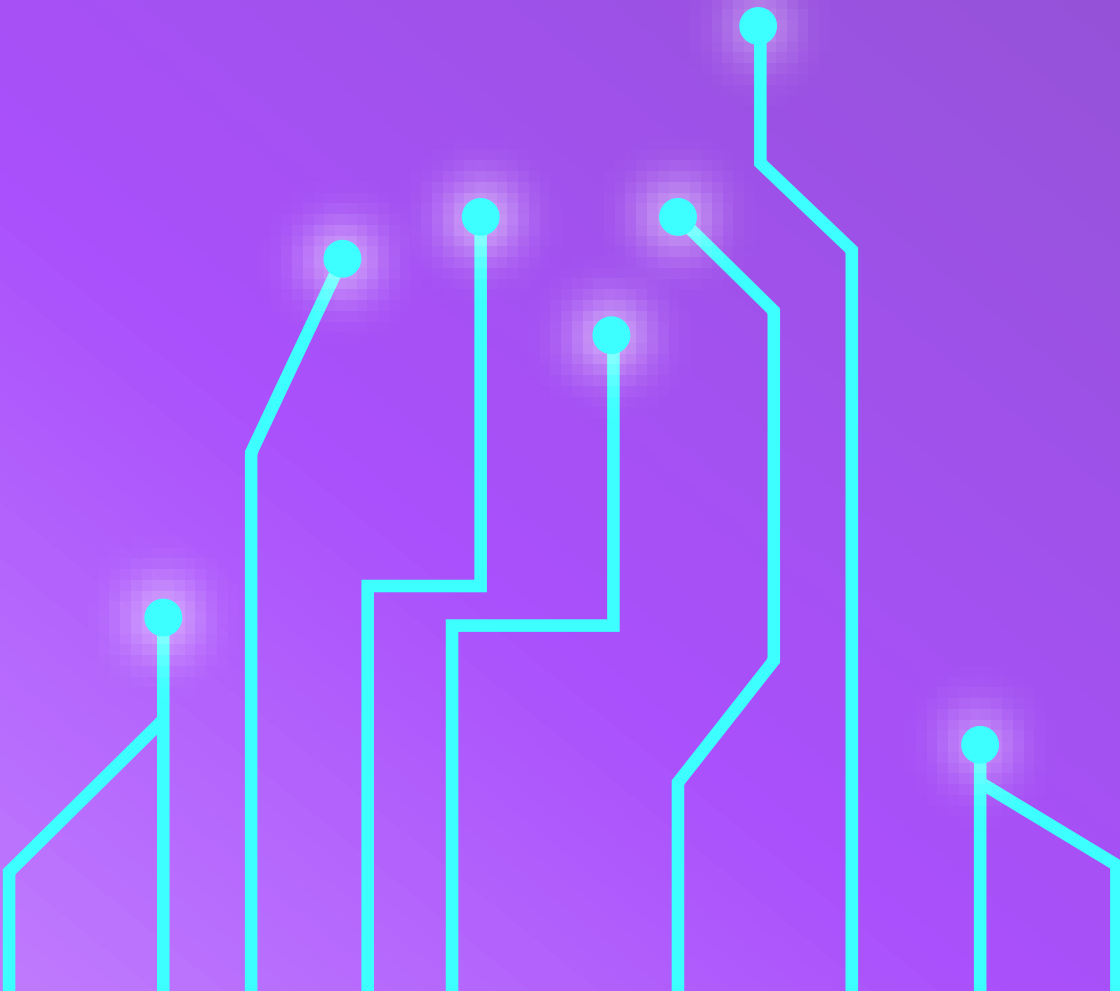




# 03. What is Algorithmic Bias?

CU1 | What is Algorithmic Bias?





### 03. What is Algorithmic Bias?

Algorithms are used to make important decisions. However, they can sometimes be biased and unfair towards certain groups of people. This is known as algorithmic bias.

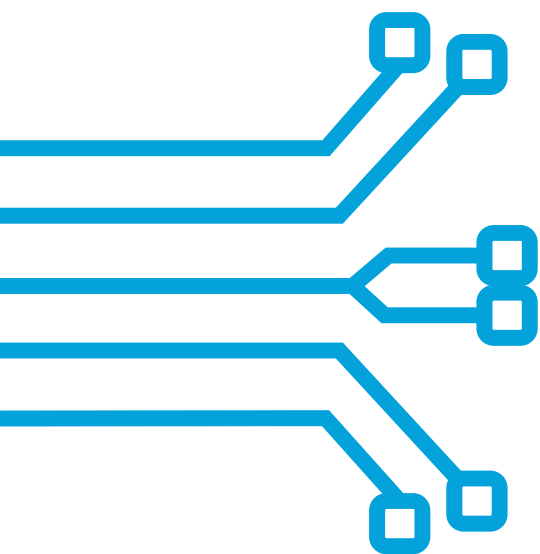
In this competence unit, students will learn about algorithmic bias, its different forms, and how to identify it. They will also explore the reasons behind bias in algorithms, including the impact of human bias on decision-making. Additionally, students will examine the potential consequences of biased algorithms on individuals and society, which can lead to discrimination and unfair treatment. By the end of this unit, students will have a better understanding of algorithmic bias and how to address it in their future work.

The knowledge outcomes for this unit include:

- **Defining Algorithmic Bias:** Students will learn about algorithmic bias and its causes, including biased data collection, skewed training data, and human decision-making. This knowledge will help them understand how bias can impact AI applications, such as facial recognition systems that misidentify certain groups.

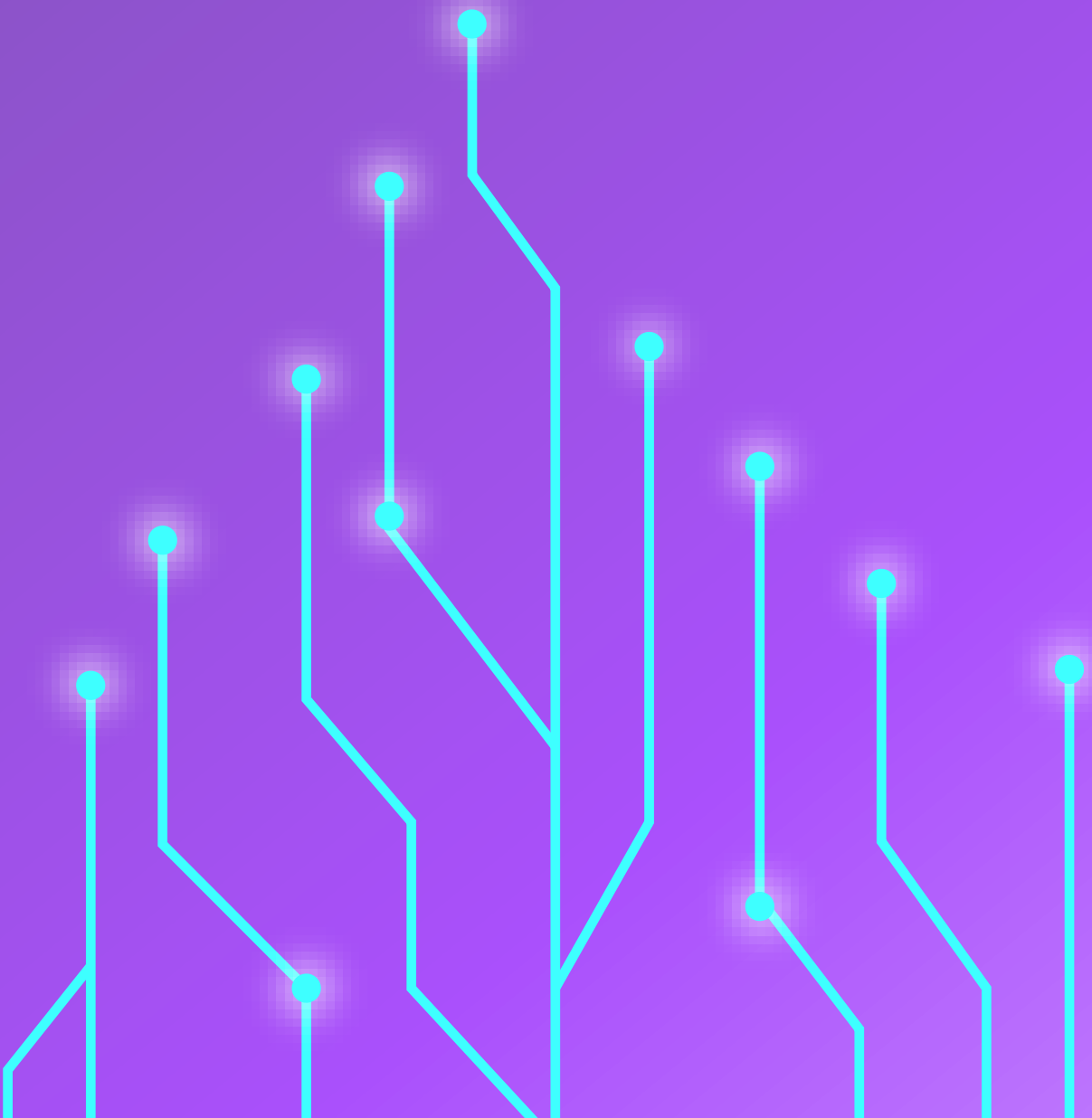


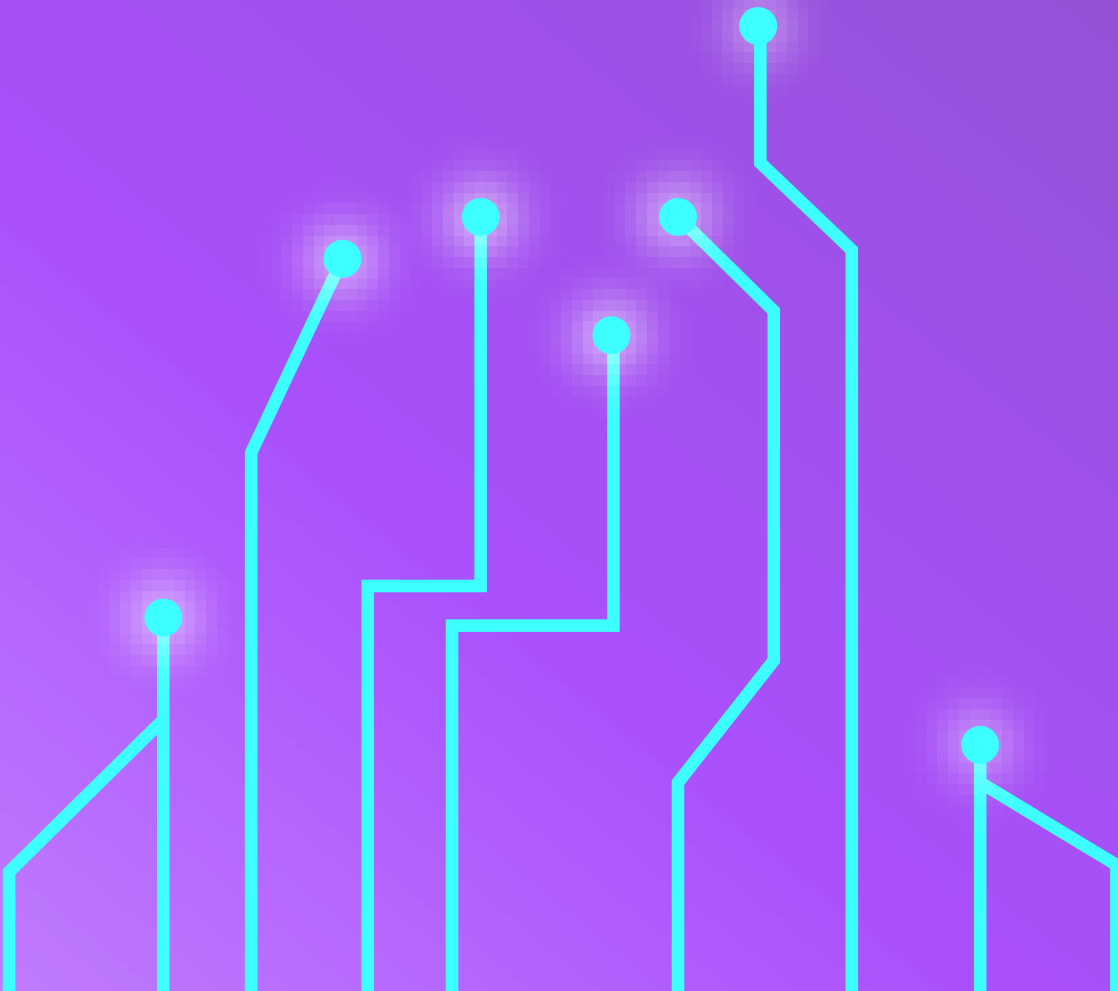
- **Identifying Types of Algorithmic Biases:** Students will learn about algorithmic biases, including data-driven, model-driven, and human-driven biases. They will understand how these biases can cause unfairness in AI systems. For instance, data-driven bias can result from unrepresentative training data, leading to biased predictions in areas like credit scoring or job applicant screening.
- **Real-World Implications of Algorithmic Bias:** In this course, students will learn about the consequences of algorithmic bias in different sectors such as healthcare, finance, and criminal justice. They will understand the need to minimise algorithmic bias in AI systems to promote fairness and equity. Examples of biased AI systems leading to negative outcomes in healthcare and criminal justice will be discussed.



# 04. Defining Algorithmic Bias

CU1 | What is Algorithmic Bias?





## 04. Defining Algorithmic Bias

Algorithmic bias is a critical aspect of AI that has garnered attention in recent years. Understanding it is essential for anyone involved in AI development, deployment, or regulation. Let's define what algorithmic bias is and why it's crucial to study.

### > What is Algorithmic Bias?

Algorithmic bias refers to systematic errors or unfairness in the outcomes of AI systems due to various factors such as biased data, flawed algorithms, or human decision-making. These biases can lead to discriminatory or unjust treatment of individuals or groups, perpetuating existing social inequalities and reinforcing stereotypes.

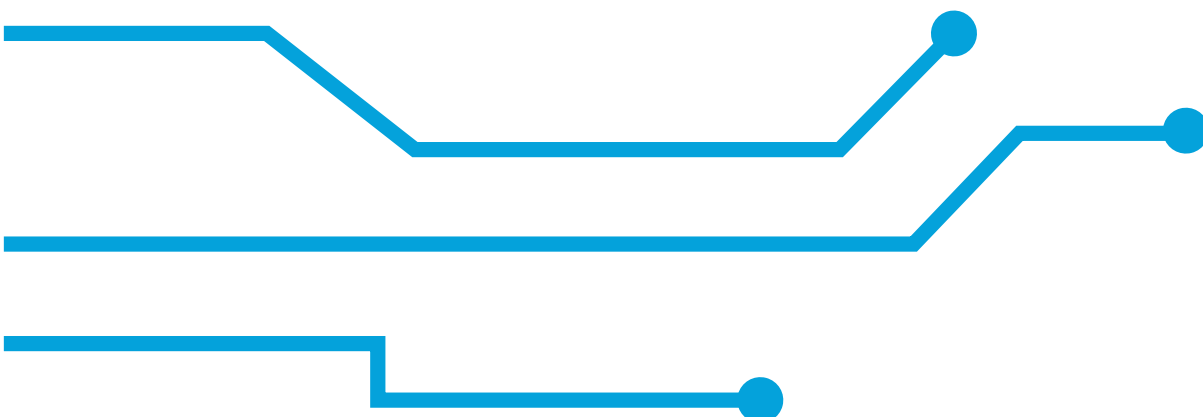
#### Why study Algorithmic Bias?

In order to explore the different forms, causes, and consequences of algorithmic bias, it's important to first understand its definition and significance. With this knowledge, we can equip ourselves with the tools to identify, mitigate, and prevent algorithmic bias in AI systems.

- 1. Ethical Implications:** Algorithmic bias can result in unfair treatment of individuals based on race, gender, age, or other protected characteristics, violating principles of fairness and equity.



2. **Social Impact:** Biased AI systems can exacerbate societal inequalities and discrimination, affecting access to opportunities, resources, and services for marginalised communities.
3. **Legal and Regulatory Concerns:** As AI technologies become more pervasive, there is growing scrutiny from lawmakers and regulatory bodies to address algorithmic bias to ensure compliance with anti-discrimination laws and protect individuals' rights.
4. **Reputation and Trust:** Organisations that deploy biased AI systems risk reputational damage and loss of public trust, which can have significant consequences for their brand image and market credibility.



## ➤ Factors contributing to biased outcomes

Several interrelated factors contribute to the emergence of biased AI systems, undermining their reliability, fairness, and effectiveness. In this section we'll explore some of the most common factors contributing to biased outcomes in AI systems.

- **Biased Data:** Biased data used to train AI systems results in algorithmic bias, which can lead to discriminatory outcomes. To mitigate this, careful considerations must be made in data collection and preprocessing, including representative sampling, bias detection and mitigation algorithms, and diverse data augmentation.
- **Flawed Algorithms:** AI systems can have biased outcomes due to flawed algorithms, design choices, model architectures, optimization procedures, or input variables. Fairness-aware machine learning, algorithmic transparency, and interpretability techniques can help mitigate such biases.
- **Human Biases:** Biases in AI systems can result from unconscious influences of developers, data scientists, and decision-makers. To avoid these biases, AI development teams should focus on diversity, ethical guidelines, and accountability mechanisms.





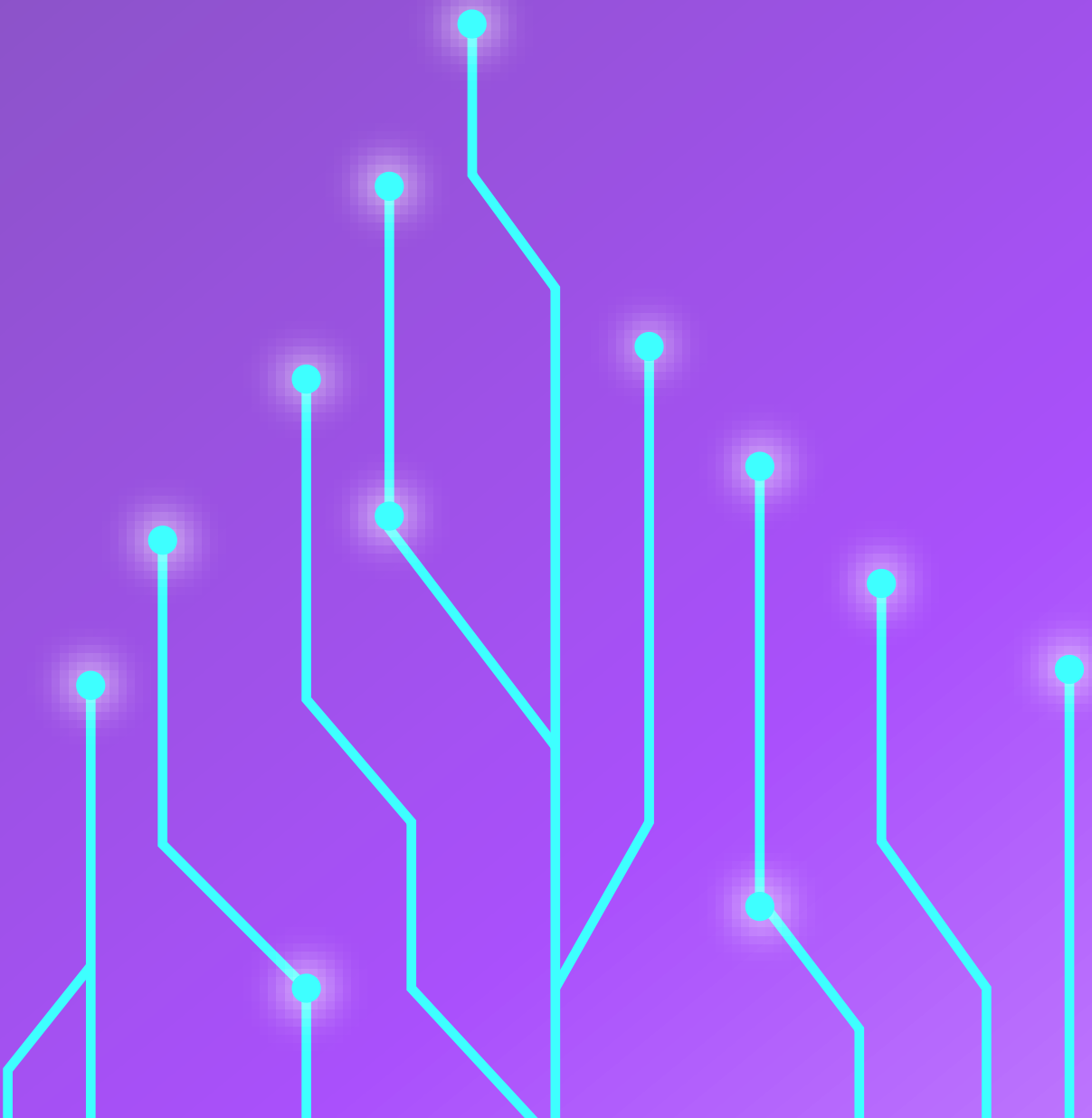
## > Examples of biased systems

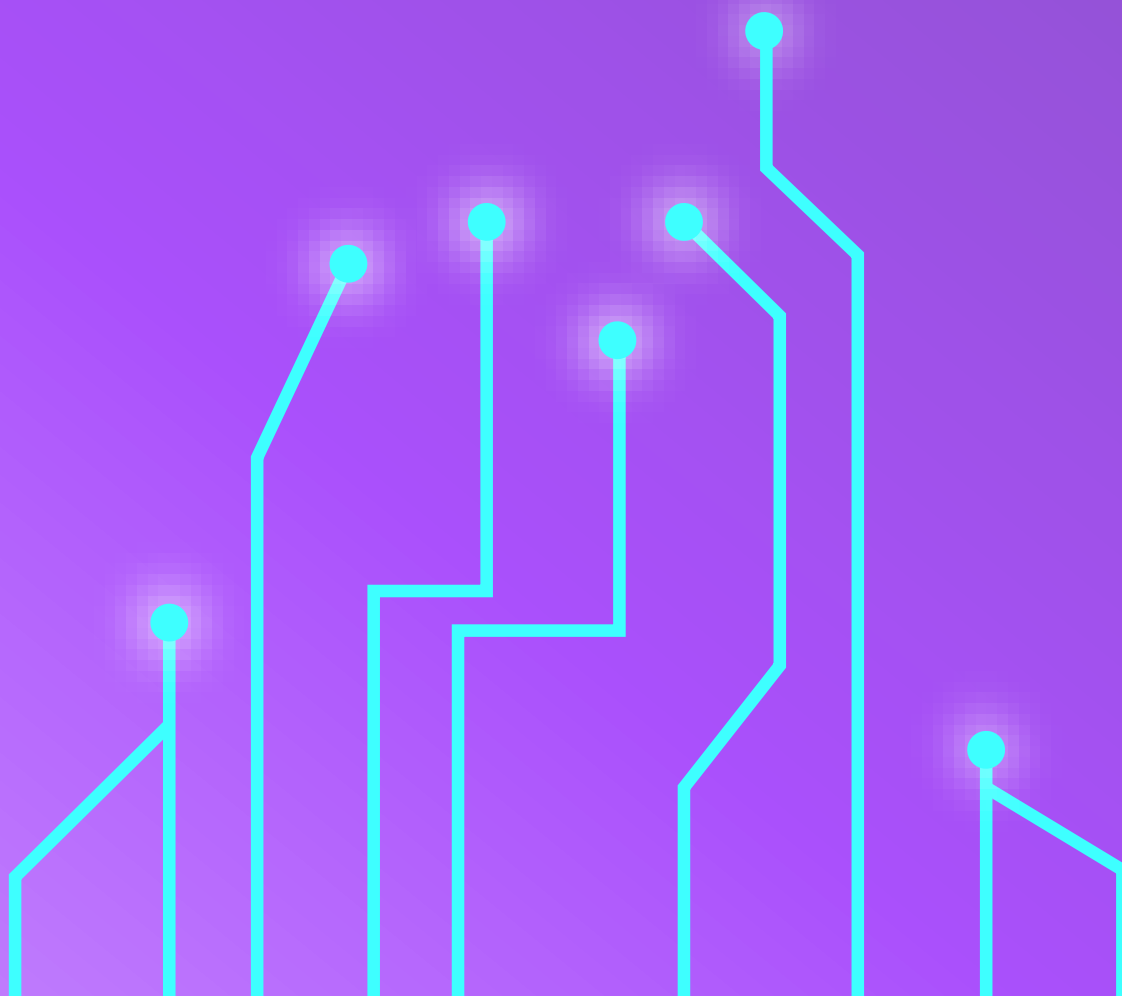
AI systems can have biases, leading to unfair outcomes. Below are some quick real-world examples of commonly biased AI systems, highlighting the potential consequences of algorithmic biases. We will explore them more profoundly in later sections of this course.

- **Facial Recognition Algorithms:** Facial recognition technology can have biases that perpetuate racial or gender disparities, leading to wrongful arrests or surveillance of specific groups. Addressing these biases is crucial for ensuring fairness and equity in AI systems and restoring public trust.
- **Predictive Policing Algorithms:** Predictive policing algorithms can perpetuate biases present in historical crime data, leading to over-policing of certain communities or demographic groups. Biased algorithms can exacerbate existing disparities in law enforcement practices and raise concerns about fairness, accountability, and potential for discriminatory outcomes in criminal justice systems.
- **Automated Hiring Systems:** Automated hiring systems may perpetuate biases, leading to discriminatory practices and limiting diversity in the workforce. Biased algorithms can learn patterns of bias from historical data, resulting in preferential treatment of certain demographic groups. Auditing and mitigating biases is crucial to ensure fairness, equity, and accountability in AI-driven recruitment processes.

# 05. Understanding Bias in AI Systems

CU1 | What is Algorithmic Bias?





## 05. Understanding Bias in AI Systems

In this section we'll explore three types of bias: **data-driven**, **model-driven**, and **human-driven**.

These biases can impact the accuracy and trustworthiness of AI systems, and understanding them is the first step in preventing them.

### > **Data-Driven Bias**

What is Data-Driven Bias?

Data-driven bias refers to biases that arise from the characteristics or distribution of the training data used to develop machine learning models.

Biased training data may reflect historical inequalities, societal prejudices, or systemic discrimination, leading to skewed representations of certain demographic groups or underrepresentation of others.

Understanding data-driven bias is essential for recognizing how biased training data can perpetuate and exacerbate existing stereotypes, inequalities, and discriminatory practices in AI systems.



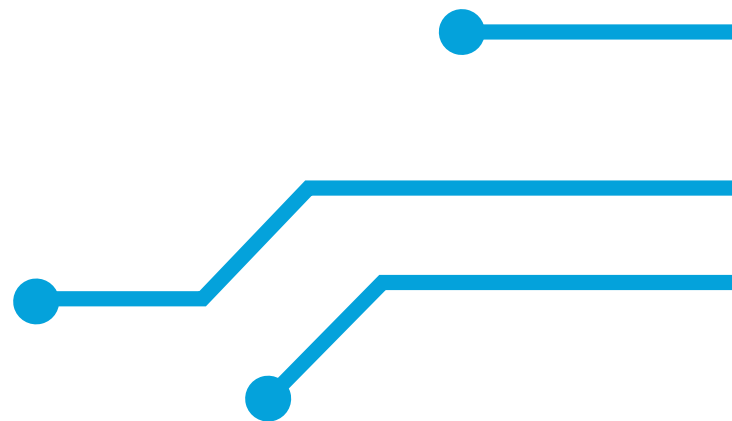
## Causes of Data-Driven Bias

- 1. Incomplete or Biased Sampling:** Training datasets may lack diversity or fail to adequately represent certain demographic groups, leading to skewed representations and biased model predictions.
- 2. Historical Biases:** Training data may reflect historical inequalities or systemic biases present in society, perpetuating discriminatory outcomes in AI systems.
- 3. Labeling Biases:** Biased or subjective labeling practices can introduce biases into training data, influencing model predictions and reinforcing existing stereotypes.



## Examples of Data-Driven Bias

- 1. Biased Facial Recognition:** Facial recognition algorithms trained on imbalanced datasets may exhibit racial or gender biases, leading to misidentification and discrimination against certain demographic groups.
- 2. Gender Bias in Language Models:** Language models trained on biased text corpora may generate gender-stereotypical or discriminatory language, reflecting and perpetuating societal biases.
- 3. Racial Bias in Predictive Policing:** Predictive policing algorithms trained on biased crime data may disproportionately target minority communities, exacerbating racial disparities in law enforcement.





## Impact of Data-Driven Bias

- 1. Reinforcement of Stereotypes:** Biased training data can reinforce existing stereotypes and prejudices, perpetuating discrimination and inequality in AI systems.
- 2. Amplification of Inequality:** Data-driven bias can exacerbate existing inequalities and disparities, leading to unfair treatment and discriminatory outcomes for marginalised groups.
- 3. Erosion of Trust:** Biased AI systems undermine trust and confidence in technology, exacerbating concerns about fairness, accountability, and transparency.

Data-driven bias is a significant challenge for fair and equitable AI systems. By understanding its causes and consequences, stakeholders can take proactive steps to mitigate bias in training data and promote inclusivity in AI.



## > Model-Driven Bias

### What is Model-Driven Bias?

Model-driven bias refers to biases that emerge from the design, structure, or optimization of machine learning models, leading to discriminatory outcomes or skewed predictions.

### Causes of Model-Driven Bias

- 1. Feature Selection Biases:** Model features selected during the modeling process may inadvertently encode biases present in the training data, leading to biased predictions or discriminatory outcomes.
- 2. Algorithmic Complexity:** Complex machine learning algorithms may capture and reinforce subtle biases present in the training data, amplifying their impact on model predictions.
- 3. Optimization Objectives:** Optimization objectives defined during the model training process may inadvertently prioritize certain outcomes over others, leading to biased or unfair predictions.





## Examples of Model-Driven Bias

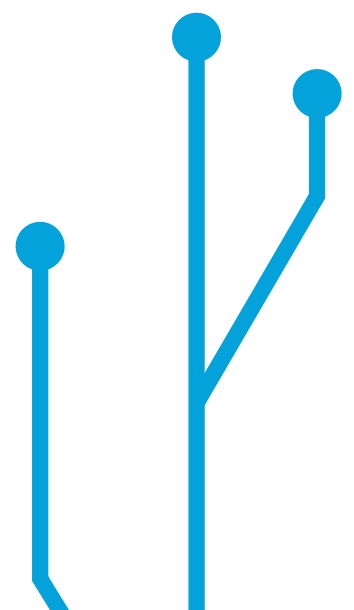
- 1. Gender Bias in Hiring Algorithms:** Automated hiring algorithms may inadvertently favor male candidates over female candidates due to biased feature selection or optimization objectives, perpetuating gender disparities in the workforce.
- 2. Racial Bias in Sentencing Algorithms:** Predictive sentencing algorithms used in criminal justice systems may disproportionately recommend harsher sentences for minority defendants, amplifying racial disparities in incarceration rates.
- 3. Socioeconomic Bias in Loan Approval Models:** Machine learning models used for loan approval may systematically deny loans to individuals from marginalized communities, exacerbating socioeconomic inequalities in access to financial services.



## Impact of Model-Driven Bias

- 1. Perpetuation of Discrimination:** Model-driven bias can perpetuate and reinforce existing discrimination and inequalities present in society, leading to unfair treatment and biased outcomes for marginalized groups.
- 2. Lack of Accountability:** Biased AI models may lack transparency and accountability, making it challenging to identify and address discriminatory practices in AI systems.
- 3. Ethical Implications:** Model-driven bias raises ethical concerns related to fairness, justice, and human rights, highlighting the need for ethical guidelines and regulations to govern AI development and deployment.

Model-driven bias poses significant challenges to the development and deployment of fair and accountable AI systems. By understanding the mechanisms and implications of model-driven bias, stakeholders can implement strategies to mitigate bias and promote fairness and equity in AI technologies.





## > Human-Driven Bias

### What is Human-Driven Bias?

Human-driven bias in AI refers to biases arising from decisions, actions, or judgments of individuals involved in development and deployment. It can stem from cognitive biases, cultural influences, and societal prejudices, leading to biased outcomes or discriminatory practices.

### Causes of Human-Driven Bias

- 1. Data Collection Biases:** Data collection biases like sampling or selection biases can lead to biased training data and skewed model predictions.
- 2. Algorithmic Design Biases:** AI algorithms can be biased due to human designer's and developer's choices, perpetuating biased outcomes in AI systems.
- 3. Interpretation and Deployment Biases:** Human interpreters and decision-makers may exhibit biases when deploying AI systems, leading to discriminatory practices and unfair treatment.

## Examples of Human-Driven Bias

- 1. Bias in Facial Recognition Systems:** Human biases in training data collection and algorithmic design can lead to racial or gender biases in facial recognition systems, resulting in misidentification or underrepresentation of certain demographic groups.
- 2. Fairness in Hiring Algorithms:** Biases in human decision-making processes, such as resume screening or interview evaluations, can perpetuate gender or racial disparities in hiring outcomes, even when using AI-based hiring algorithms.

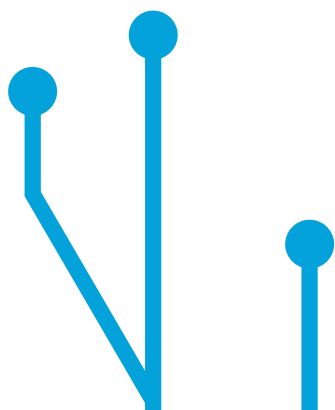




## Impact of Human-Driven Bias

- 1. Exacerbation of Existing Inequalities:** Human-driven biases in AI can exacerbate existing inequalities and disparities in society. Biased data collection, algorithmic design, and interpretation can lead to unfair treatment for marginalized groups, perpetuating discrimination and hindering social progress.
- 2. Erosion of Trust and Public Confidence:** AI systems tainted by human biases can erode public trust and confidence in technology. Concerns about fairness, transparency, and accountability can arise, hindering the adoption and acceptance of AI in various sectors.
- 3. Reduced Effectiveness of AI Systems:** Human-driven biases can undermine the effectiveness of AI systems. Biased training data or biased interpretations by humans can lead to inaccurate predictions, flawed recommendations, and suboptimal outcomes, hindering the potential benefits of AI.

Human bias is a significant challenge for creating fair and accountable AI systems. By understanding and mitigating algorithmic biases, stakeholders can build more trustworthy and transparent AI systems.





# Charlie



Co-funded by  
the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.



Universitat  
de les Illes Balears



ENGAGING PEOPLE



INNOVATION TRAINING CENTER



AARHUS UNIVERSITY



VAMK UNIVERSITY OF APPLIED SCIENCES

helixconnect



2022-1-ES01-KA220-HED-000085257