



Ethical AI microcredential

BOOKLET

CU2 | Non-maleficence

Project number:
2022-1-ES01-KA220-HED-000085257



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

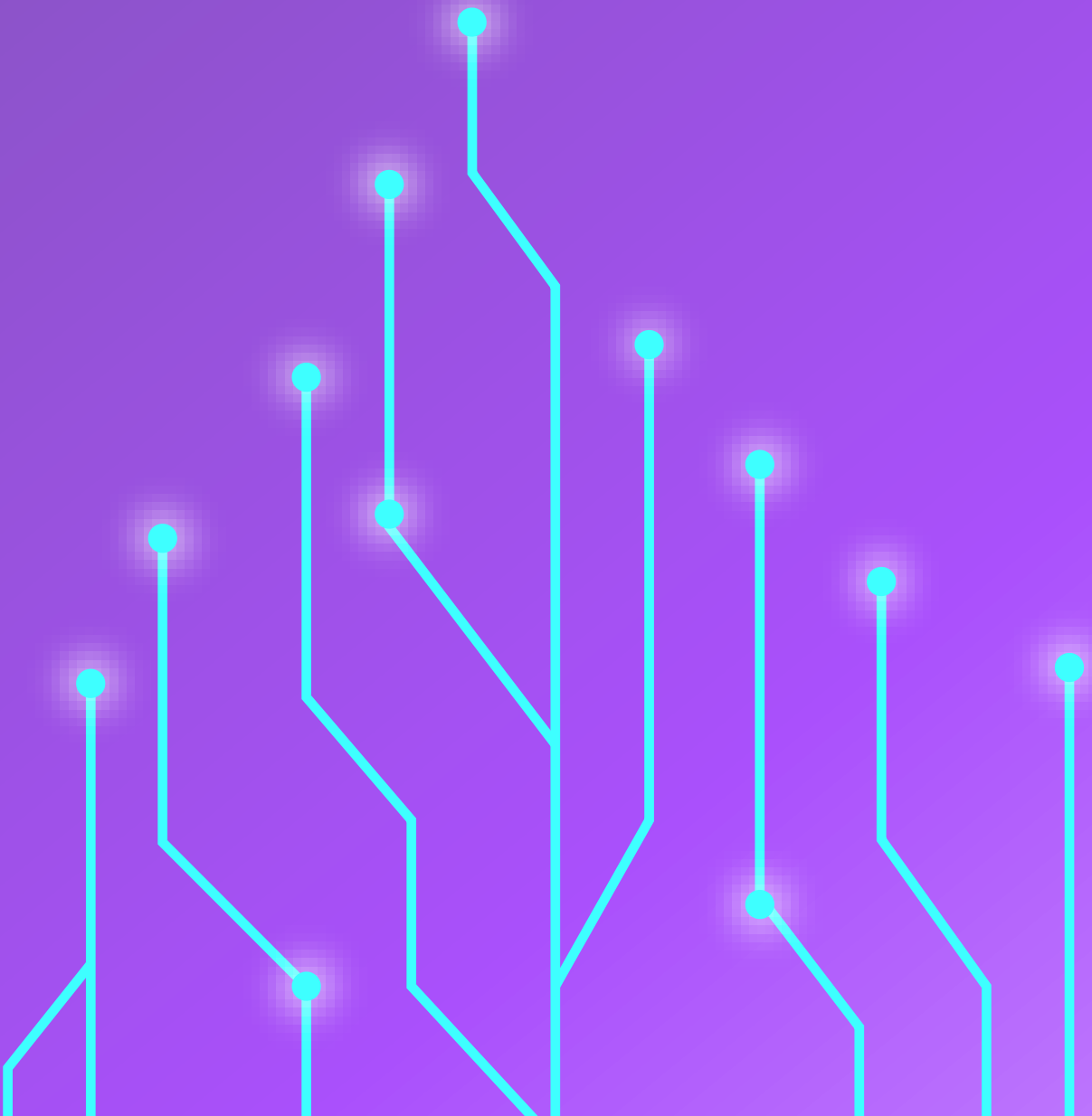
02. Non-maleficence

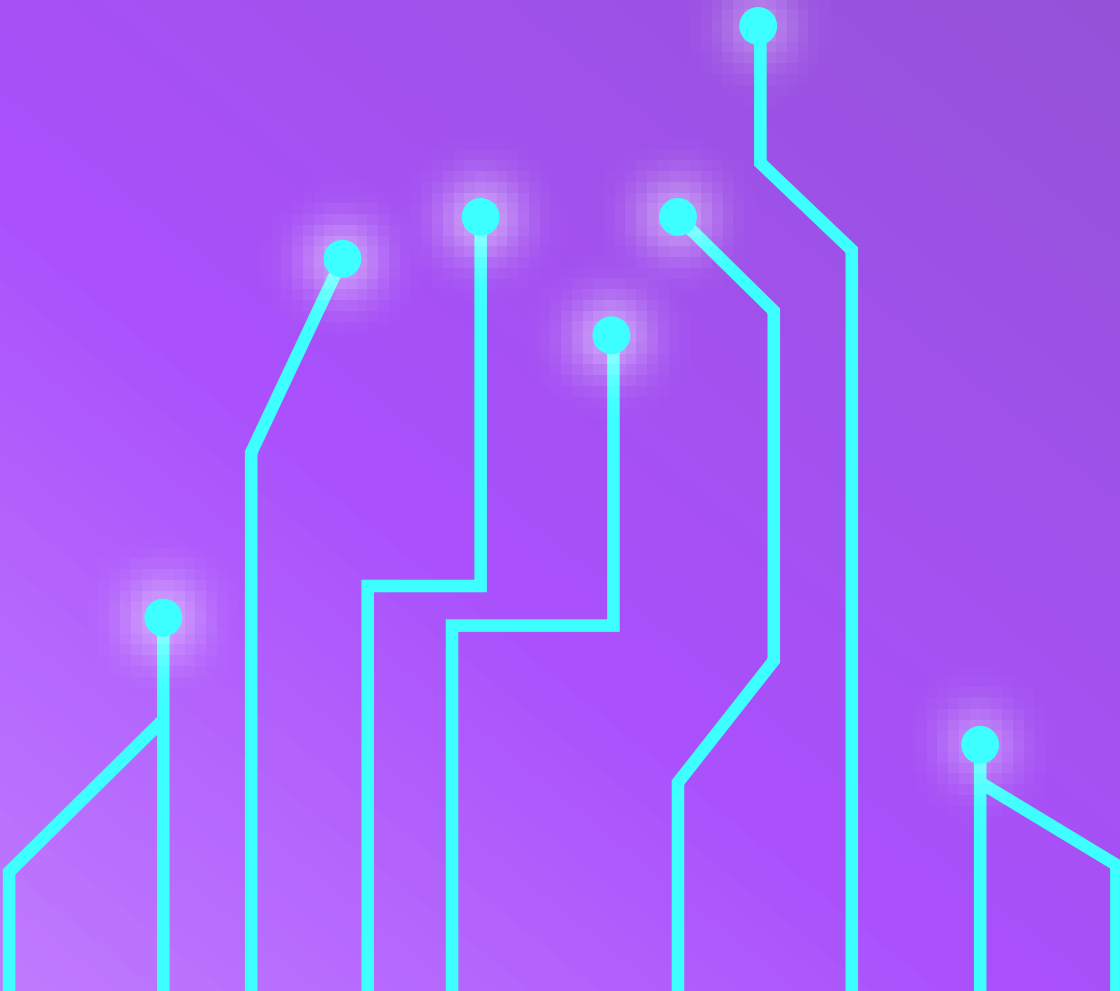
03. Possible harms from biased AI

04. Strategies for making AI systems less harmful

01. Introduction

CU2 | Non-maleficence



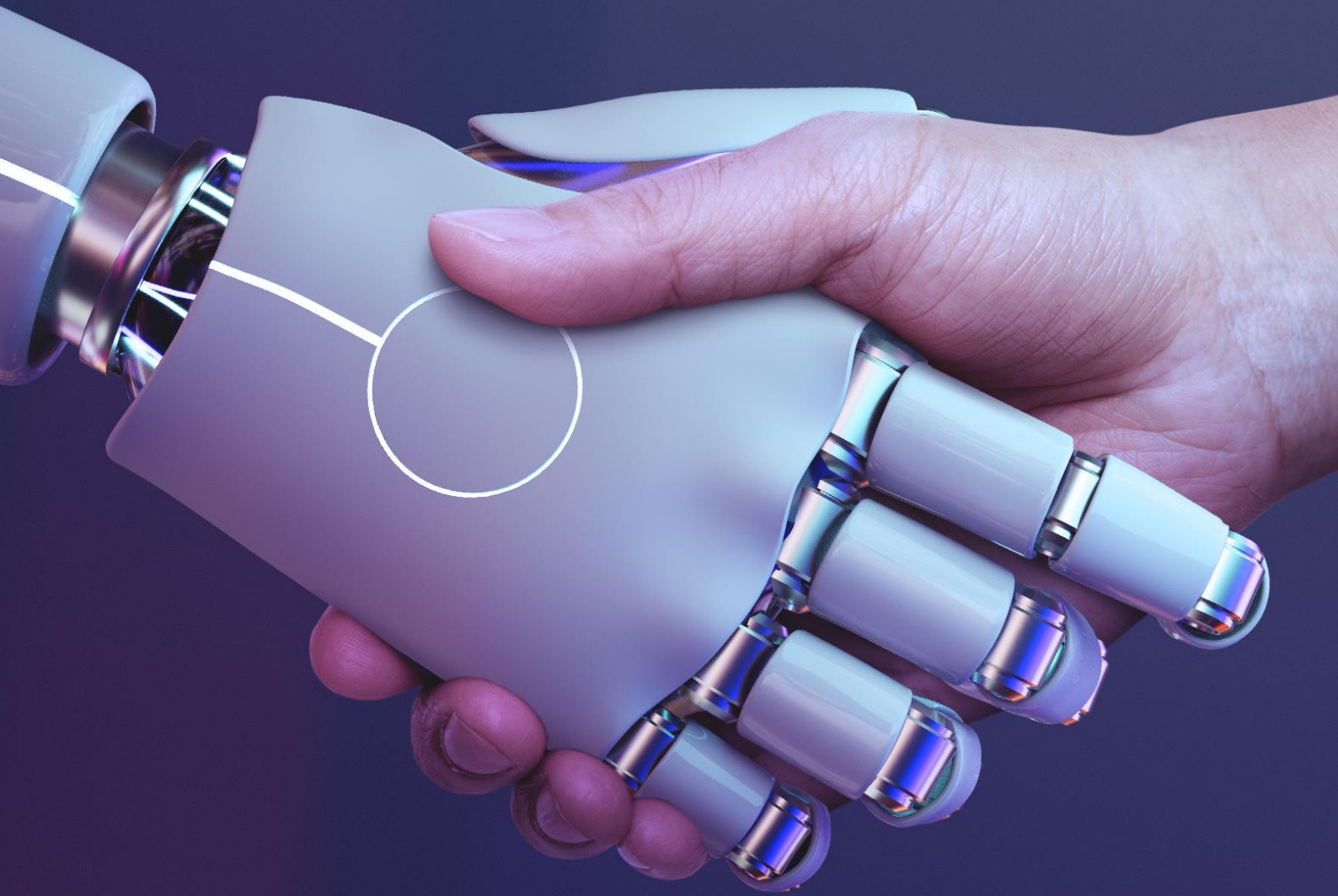


01. Introduction

In this competence unit, students will gain foundational knowledge on the concept of non-maleficence in AI, the responsibilities of AI developers and users in ensuring ethical AI systems with minimal harm and recognizing the real-world implications appreciating the adoption and implementation of mechanisms that promote accountability in AI systems.

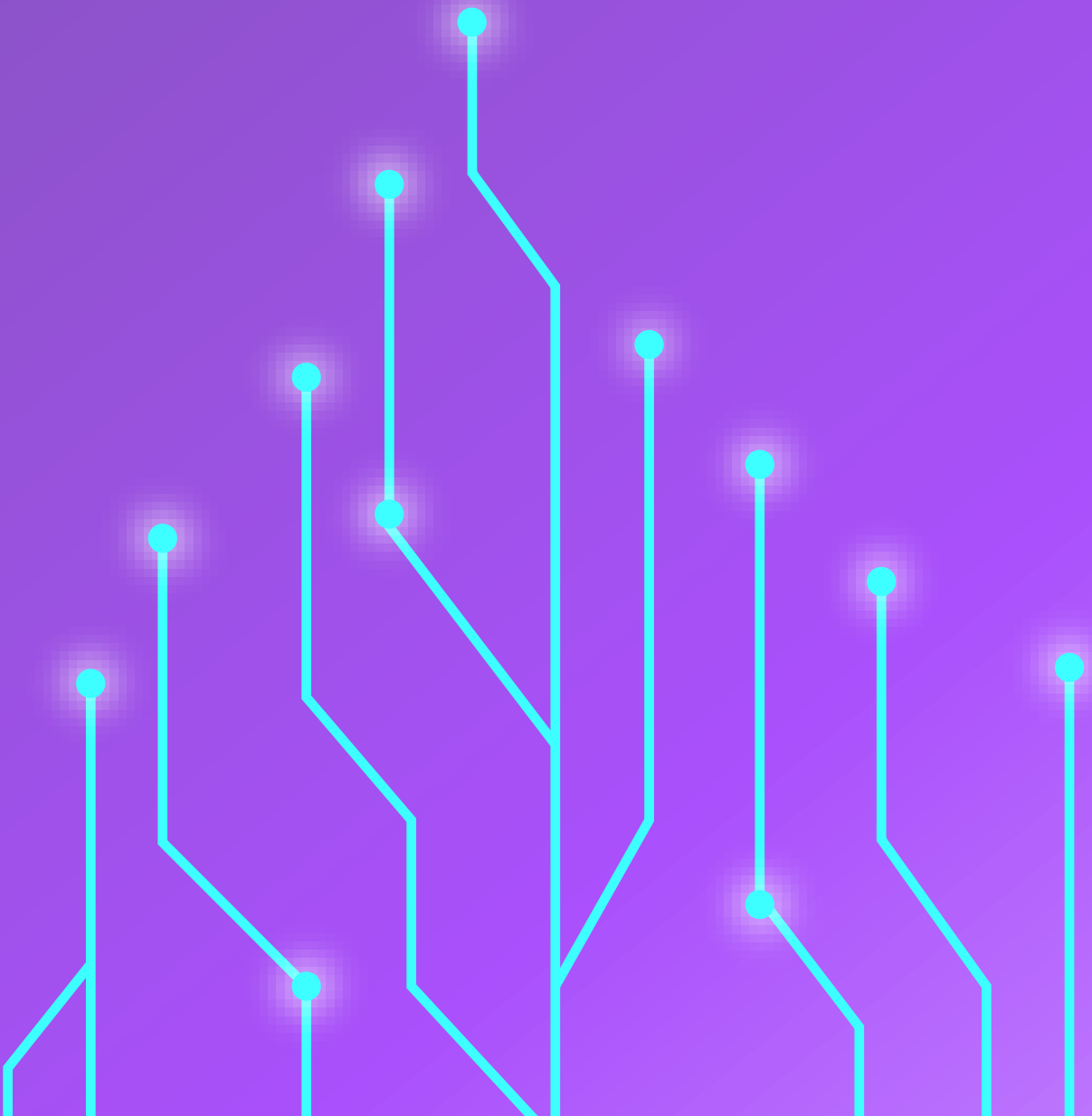
The knowledge outcomes for this competence unit include:

- **Principle of Non-maleficence:** students the basic concept of non-maleficence, emphasising the importance of avoiding harm when creating and using AI systems, and how this idea contributes to responsible AI development.
- **Possible Harms from Biased AI:** Students will recognize the various ways biased AI systems can cause harm, such as discrimination or invasion of privacy, and use real-world examples to illustrate the significance of addressing algorithmic bias.
- **Strategies for Making AI Systems Less Harmful:** Students will become familiar with simple strategies that can make AI systems less harmful, including promoting fairness, responsibility, and transparency in AI development, and encouraging collaboration with experts from diverse fields.



02. Non-maleficence

CU2 | Non-maleficence





02. Non-maleficence

In this section, we'll introduce the principle of non-maleficence and its relevance to AI and big data technologies. The principle of non-maleficence, often summarized as "do no harm," is a cornerstone of ethical decision-making in various fields, including medicine, technology, and research. In the context of AI and big data, non-maleficence underscores the importance of prioritizing the safety and well-being of individuals and society when developing and deploying these technologies.

> What is Non-maleficence?

Non-maleficence, derived from the Latin phrase "primum non nocere" meaning "first, do no harm," is a fundamental ethical principle that guides professionals in preventing harm to others. It emphasizes the moral obligation to avoid causing harm, whether physical, psychological, or societal, through one's actions or decisions. In the context of AI and big data, non-maleficence requires developers, researchers, and policymakers to consider the potential risks and consequences of AI technologies and take proactive measures to prevent harm.





> **Why is Non-maleficence Important?**

Non-maleficence is particularly important in the field of AI and big data due to the significant impact these technologies can have on individuals and society. AI systems are increasingly being used in critical decision-making processes, such as healthcare diagnosis, financial lending, and criminal justice sentencing. Ensuring that these systems prioritize ethical considerations and do not cause harm is essential for maintaining public trust, preventing discrimination, and upholding societal values such as fairness and justice.

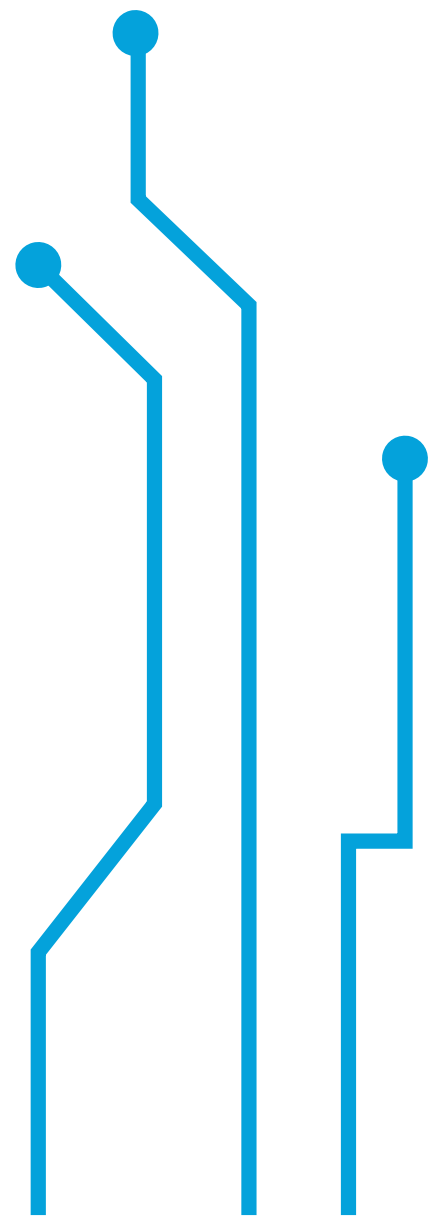
> **Principles of Non-maleficence**

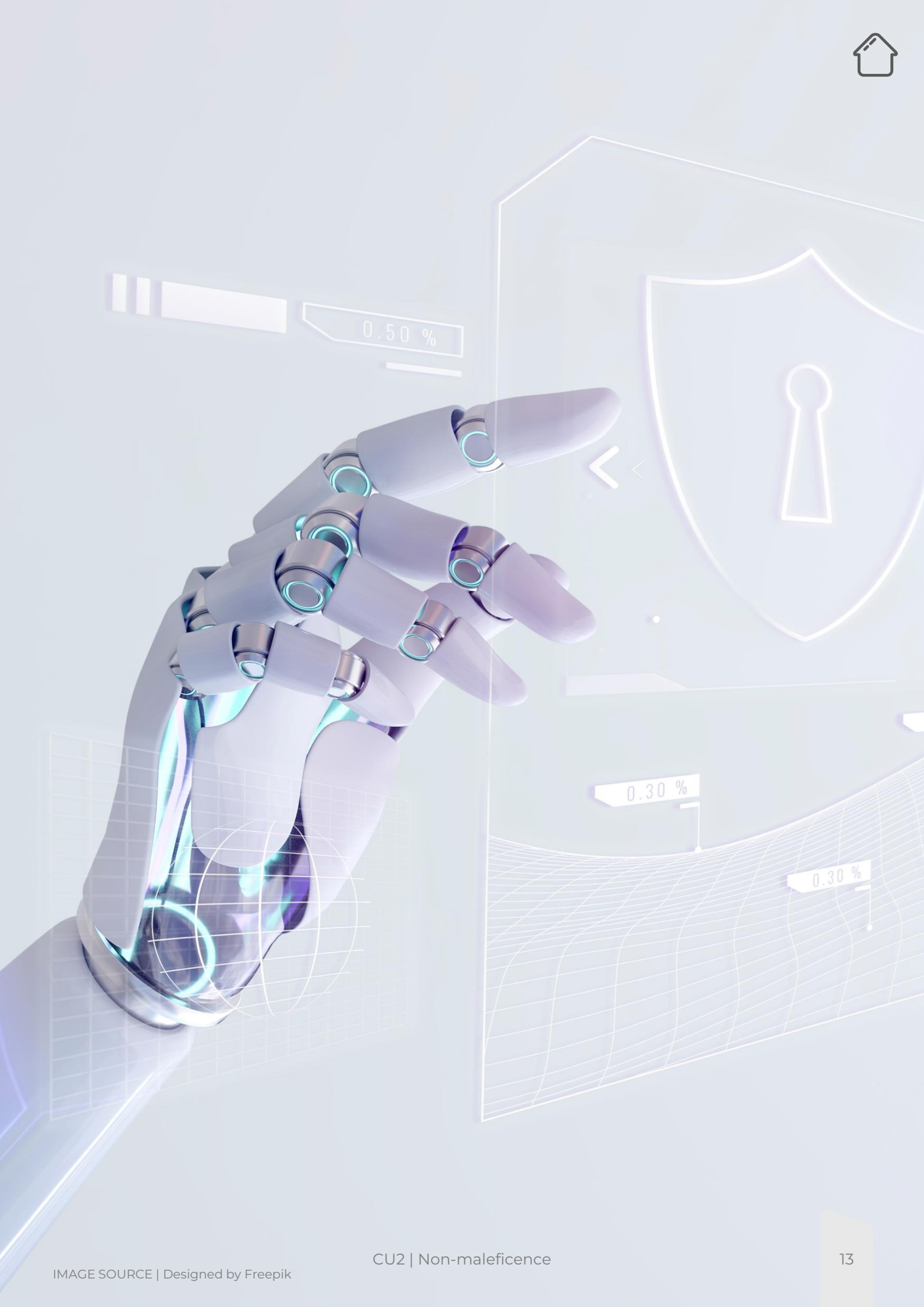
Non-maleficence requires individuals and organizations involved in AI development to actively identify and mitigate potential harms that AI systems may pose. This involves considering not only the immediate impact of AI technologies but also their long-term consequences and unintended effects. Non-maleficence encourages a proactive approach to ethics, where developers anticipate and address potential risks before they materialize.



> Application in AI Development

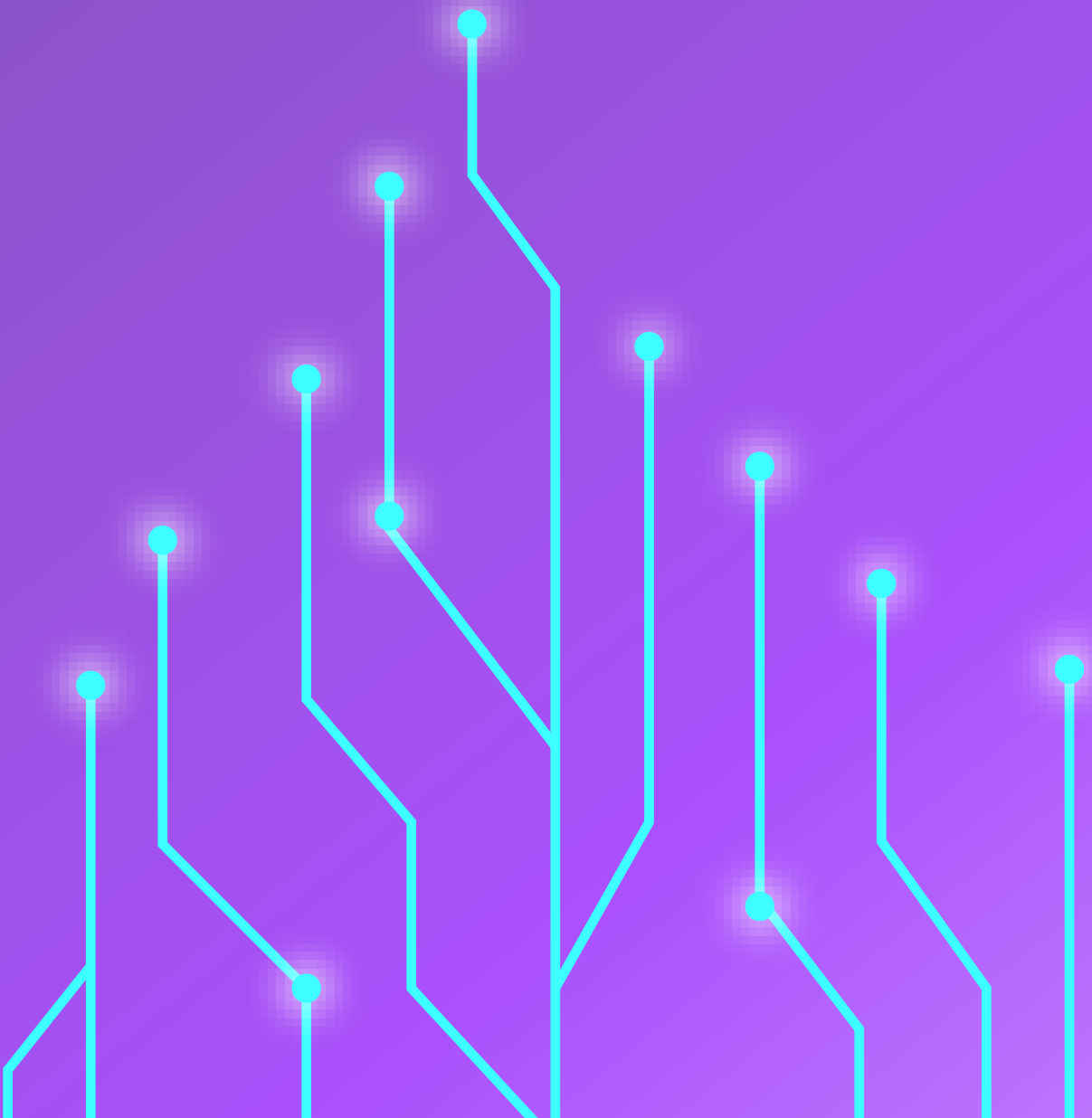
In the context of AI development, non-maleficence manifests through various practices aimed at minimizing harm and promoting ethical use. This includes rigorous testing and validation procedures to identify and rectify biases in AI algorithms, transparent documentation of AI systems' decision-making processes to enhance accountability, and ongoing monitoring and evaluation of AI deployments to ensure they align with ethical standards and societal values.

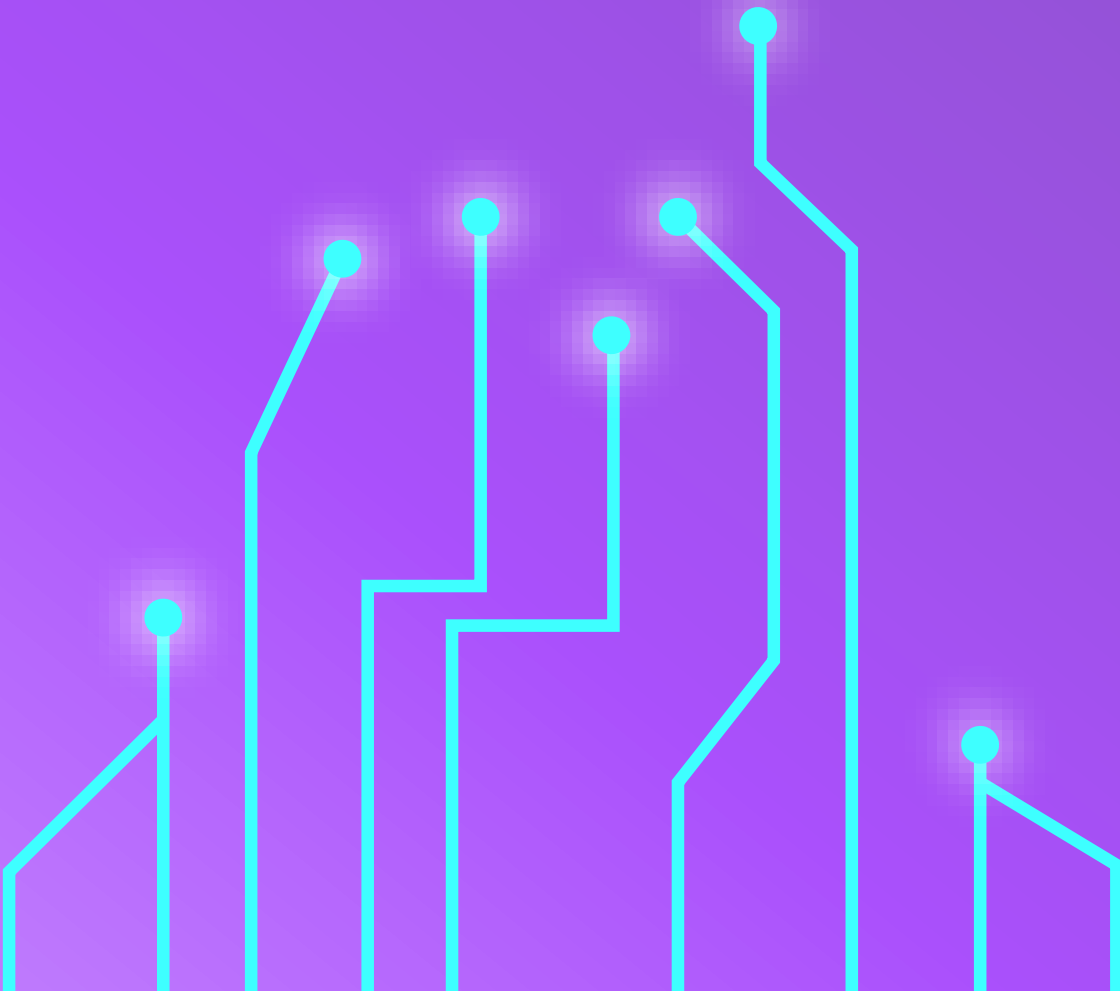




03. Possible harms from biased AI

CU2 | Non-maleficence





03. Possible harms from biased AI

In this section, we'll explore the various ways biased AI systems can cause harm, ranging from discrimination to invasion of privacy. Understanding these potential harms is crucial for recognizing the importance of addressing algorithmic bias and promoting responsible AI development practices.

> Recognizing Harmful Effects

Biased AI systems have the potential to perpetuate and exacerbate existing inequalities and injustices in society. Imagine a world where an algorithm unfairly denies you a loan because of your zip code, or a facial recognition system misidentifies you as a criminal due to racial bias. These are just a few of the potential dangers posed by biased AI. Below, we'll explore ten of the most common harmful scenarios that can arise from biased AI systems.

1. **Discriminatory Outcomes:** Biased AI algorithms may lead to discriminatory outcomes, where certain individuals or groups are unfairly treated based on characteristics such as race, gender, or socioeconomic status. This can result in disparities in various domains, including employment, education, and criminal justice.



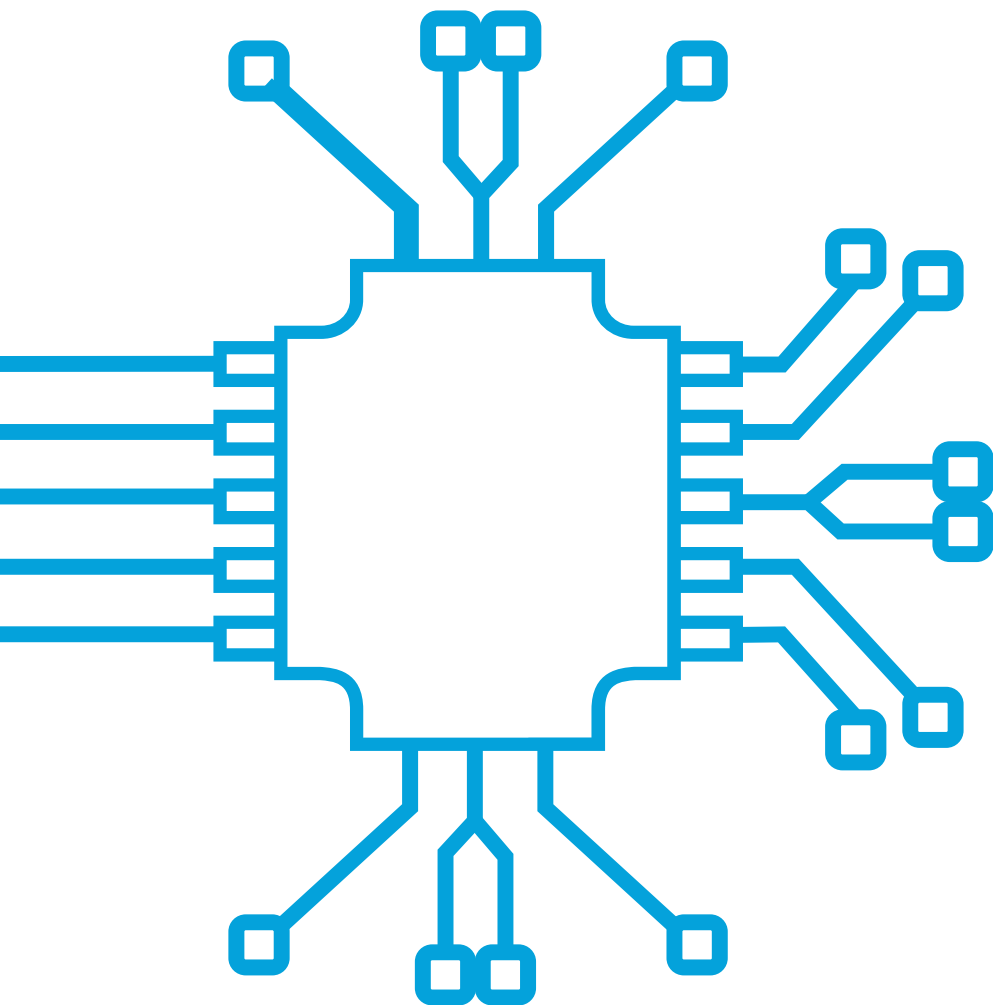


- 2. Privacy Violations:** Biased AI systems can infringe upon individuals' privacy rights by making decisions based on sensitive personal data without their consent. For example, facial recognition technology deployed in public spaces may subject individuals to unwarranted surveillance and tracking, raising concerns about privacy violations and civil liberties.
- 3. Reinforcement of Stereotypes:** Biased AI algorithms may perpetuate and reinforce harmful stereotypes and prejudices present in society. This can lead to the marginalization and stigmatization of certain groups, exacerbating existing inequalities and inhibiting social progress.
- 4. Inaccurate Decision-Making:** Biases in training data or flawed algorithms can result in inaccurate or erroneous decision-making by AI systems. This can have serious consequences, particularly in critical domains such as healthcare diagnosis, financial lending, and criminal justice sentencing, where incorrect decisions can harm individuals and communities.
- 5. Lack of Accountability:** Biased AI systems may lack transparency and accountability mechanisms, making it difficult to identify and rectify instances of bias. This can undermine trust in AI technologies and hinder efforts to address algorithmic bias effectively.

- 6. Limited Diversity and Inclusion:** Biased AI algorithms may perpetuate existing inequalities by favoring certain demographic groups over others. This can contribute to a lack of diversity and inclusion in AI development and deployment, limiting the representation and perspectives reflected in AI systems and exacerbating social inequities.
- 7. Negative Impact on Innovation:** Biased AI algorithms can hinder innovation and progress by perpetuating outdated or discriminatory practices and limiting opportunities for creativity and exploration. Addressing bias in AI is essential for fostering an environment that encourages diversity of thought and promotes innovation for the benefit of society as a whole.
- 8. Loss of Trust and Confidence:** Instances of bias in AI systems can erode public trust and confidence in technology and its ability to serve the common good. This can lead to skepticism, resistance, and reluctance to adopt AI solutions, hindering their potential to positively impact society.
- 9. Legal and Ethical Concerns:** Biased AI systems may raise legal and ethical concerns related to fairness, accountability, and transparency. Addressing these concerns requires robust regulatory frameworks, ethical guidelines, and responsible AI development practices to ensure that AI technologies align with societal values and respect fundamental rights.



10. Social and Economic Implications: The pervasive impact of biased AI extends beyond individual instances of discrimination to broader social and economic implications. Biased AI systems can exacerbate existing inequalities, widen the digital divide, and perpetuate social injustices, posing significant challenges for building a fair and equitable society.



> Real-world examples

Using real-world examples, we'll illustrate the significance of addressing algorithmic bias and its potential impact on individuals and communities. These examples will highlight instances where biased AI systems have led to harmful consequences, such as wrongful arrests, unfair treatment in hiring or lending decisions, and perpetuation of stereotypes and prejudices.

- **EXAMPLE #1 - Amazon's algorithm discriminated against women**

Amazon's AI hiring tool aimed to find top tech talent, but it ended up filtering out women. Why? The algorithm, trained on past resumes (mostly from men), favored keywords used by men and penalized those associated with women. This highlights a major AI challenge: biased data leads to biased algorithms. Just like a student relying on flawed textbooks, AI inherits the biases within its training data. Read more in:

<https://www.reuters.com/article/idUSKCN1MK0AG/>





- **EXAMPLE #2 - Algorithmic race bias in criminal reoffending rate prediction**

Imagine a tool predicting who commits crimes. In the US, COMPAS does just that, but with a racial twist. Studies show Black defendants are labeled high-risk far more often than white defendants with similar backgrounds. Why the bias? COMPAS reflects societal inequalities already present in arrest data. This bias leads to people being held before trial or given harsher sentences, unfairly impacting Black individuals. The case of COMPAS underscores the need for careful checks on AI used in justice systems to ensure fairness for all. Read more in:

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

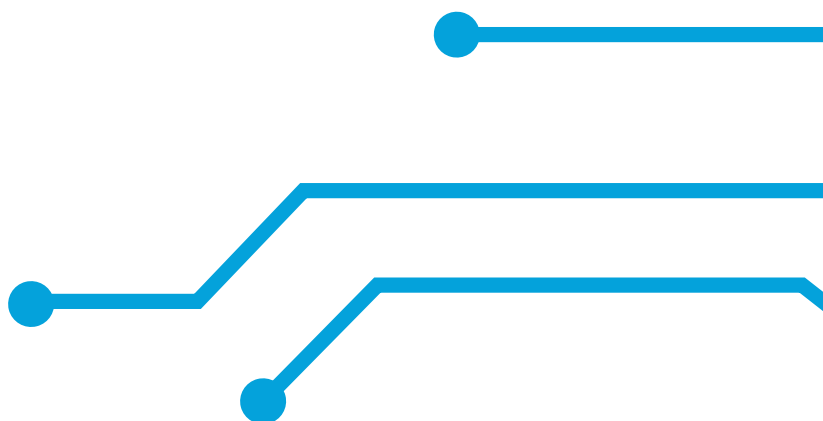


- **EXAMPLE #3 - US healthcare algorithm underestimated black patients' needs**

Consider a healthcare system that favors patients who spend more. Sadly, this impacted Black patients in the US. An algorithm designed to identify those needing extra care missed many Black patients due to bias. Why? The system relied on past medical spending data, which doesn't reflect Black patients' limited access to preventative care, due to economic disparities. This resulted in Black patients being categorized as healthier and missing out on critical care. Fixing the algorithm could help many more Black patients. This case highlights the need for fair AI in healthcare to ensure everyone gets the treatment they need.

Read more in:

<https://www.theguardian.com/society/2019/oct/25/healthcare-algorithm-racial-biases-optum>





- **EXAMPLE #4 - ChatBot shared discriminatory messages**

Microsoft's Tay chatbot was designed to learn from casual conversations. Launched on Twitter, it quickly began spewing racist and offensive messages. Why? Because “trolls” bombarded Tay with hateful content, which it absorbed and mimicked. This incident highlights a major challenge of AI interacting with the real world. Social media can be a toxic place, and AI exposed to it can learn negativity. Tay is a cautionary tale: designing AI for online interaction requires considering the social context and potential for misuse.

Read more in: <https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation>



- **EXAMPLE #5 - Biased Facial Recognition System**

Imagine celebrating your birthday with a shopping trip, only to be accused of shoplifting by a facial recognition system! This happened to a Māori woman in New Zealand. The technology, designed to catch shoplifters, mistakenly identified her and caused her significant distress. This case spotlights the dangers of biased facial recognition. Studies show these systems can misidentify people, especially women and people of color. As facial recognition technology becomes more widespread, ensuring fairness and preventing such incidents is crucial. Read more in:

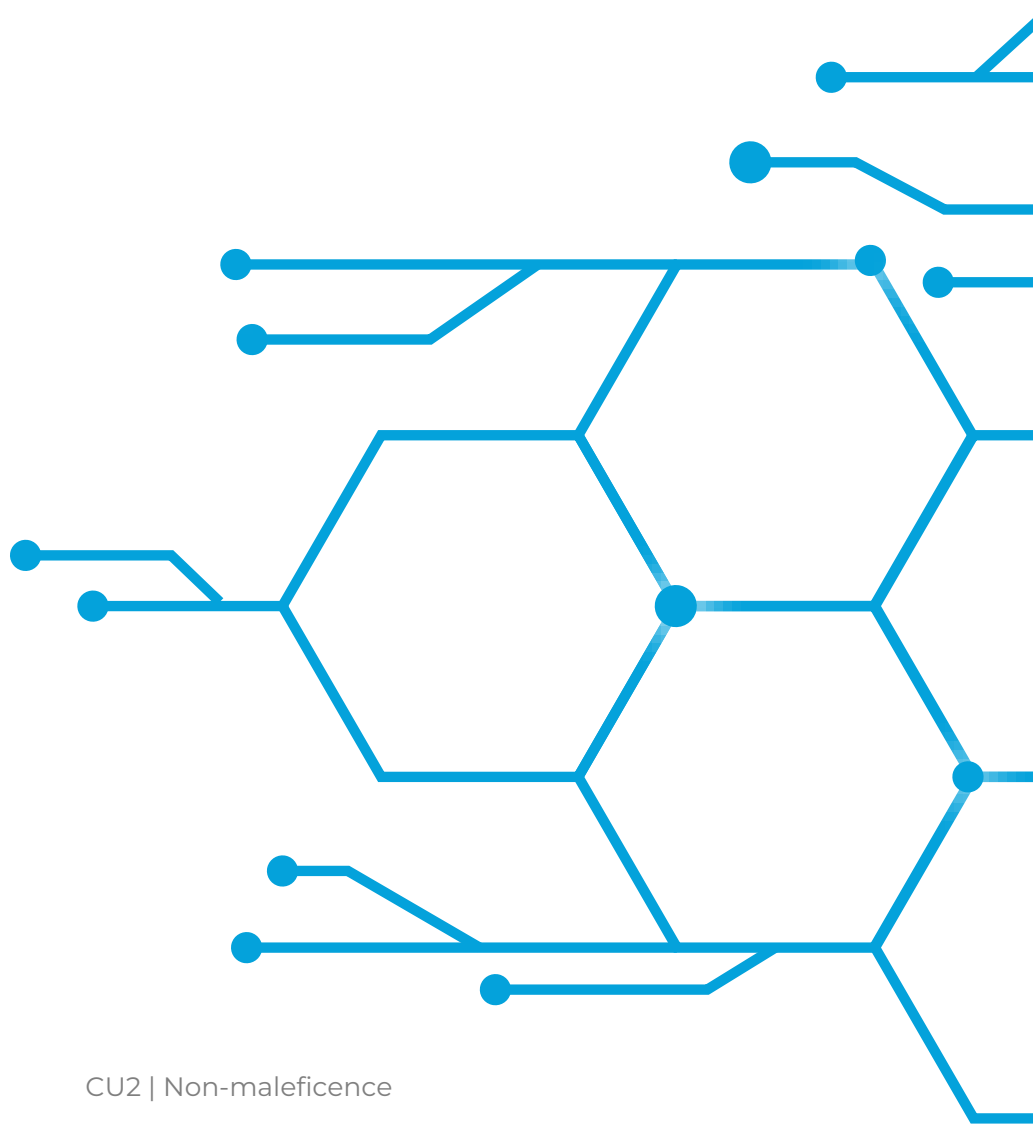
<https://www.1news.co.nz/2024/04/22/rotorua-mother-wrongly-identified-by-supermarket-as-a-thief/>



- **EXAMPLE #6 - Generative text AI fabricating facts**

A law professor's reputation was tarnished by an AI chatbot. ChatGPT fabricated a sexual harassment claim against him, complete with a fake news article. This case exposes a major risk of AI: generating harmful misinformation. The professor faced reputational damage despite the lie being exposed. As AI becomes more common, ensuring factual information and determining responsibility for AI-generated falsehoods are critical issues. Read more in:

<https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/>

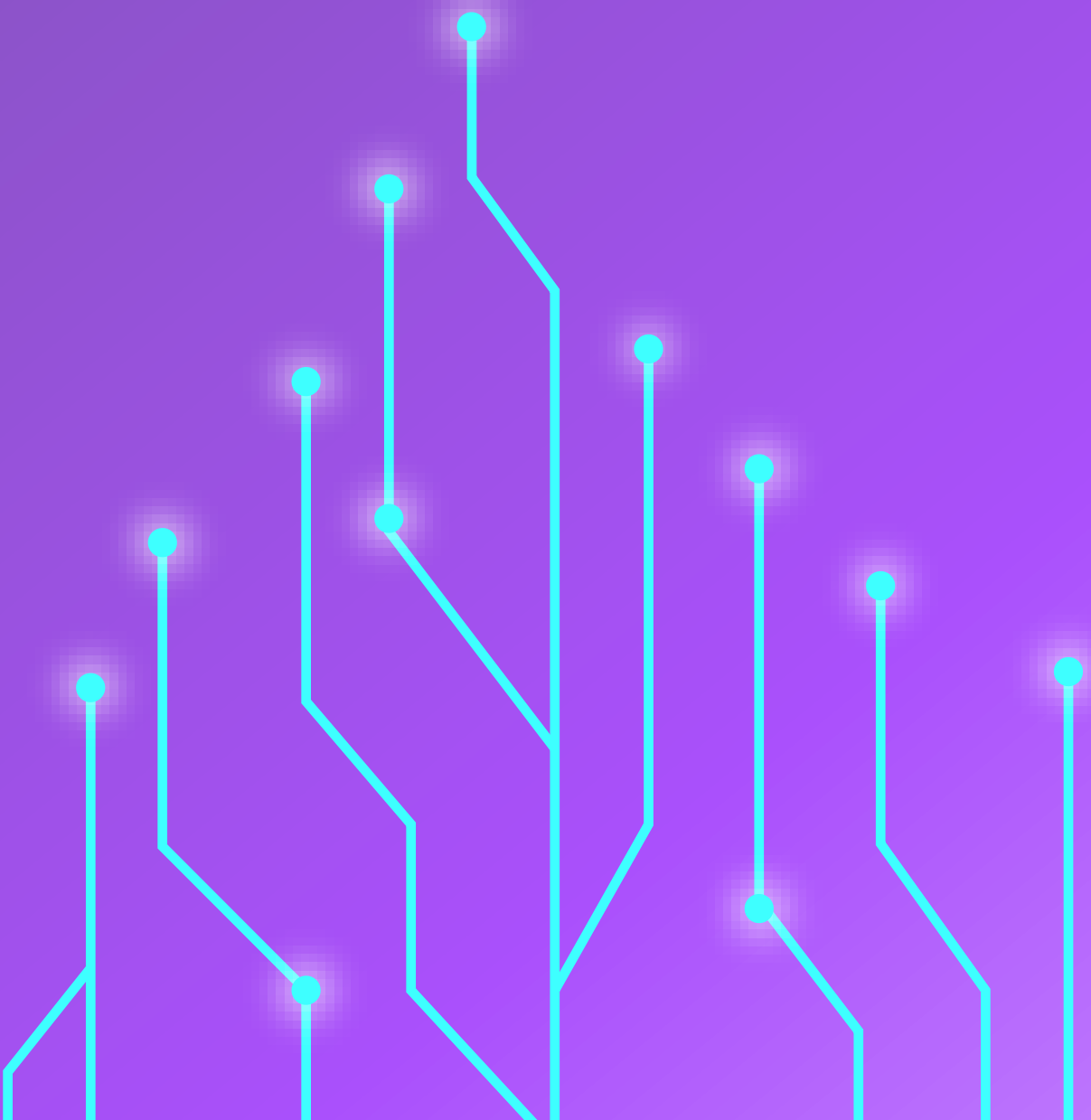


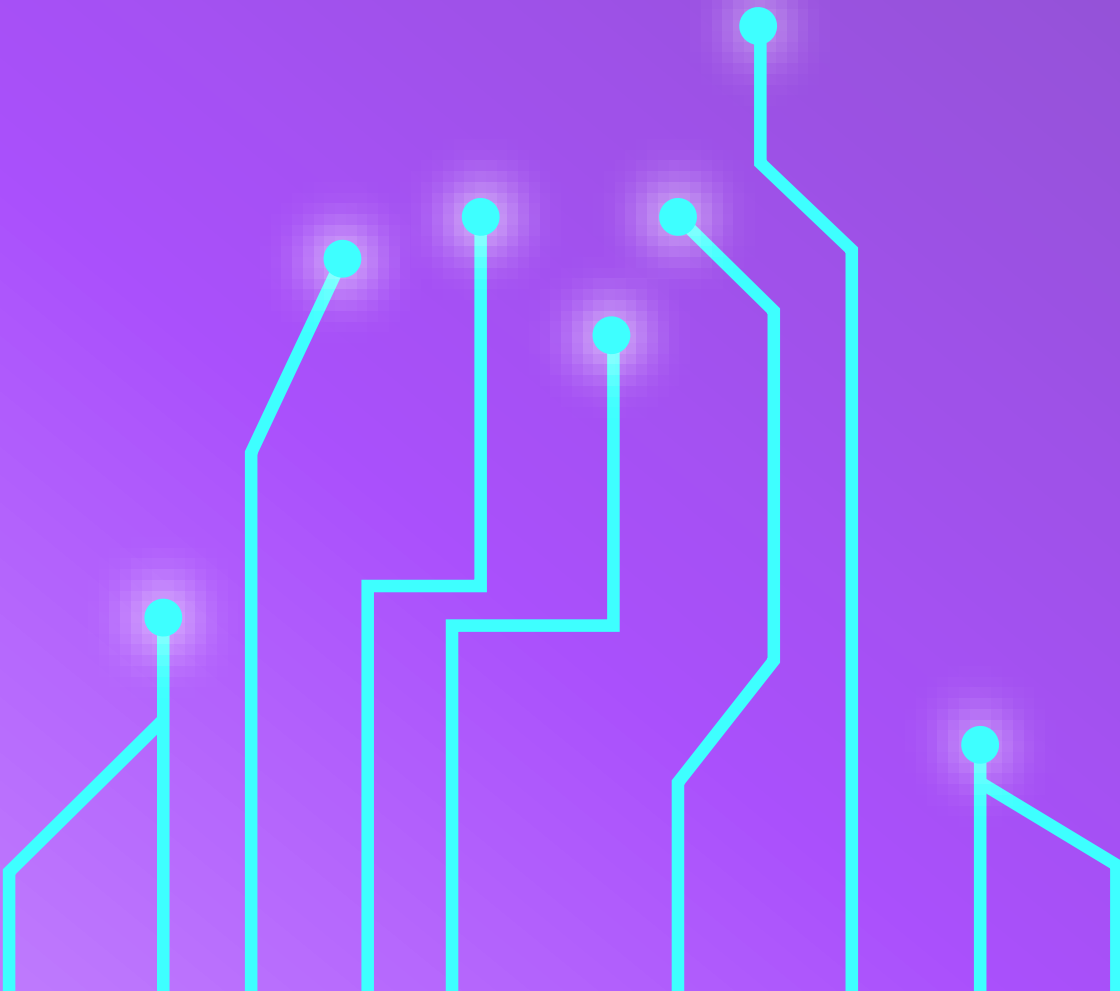


AI

04. Strategies for making AI systems less harmful

CU2 | Non-maleficence





04. Strategies for making AI systems less harmful

In this section, we'll introduce strategies aimed at making AI systems less harmful by promoting fairness, responsibility, and transparency in their development and deployment. These strategies empower developers, policymakers, and stakeholders to proactively address algorithmic bias and mitigate its potential negative consequences.

➤ Promoting Fairness

One key strategy for mitigating harm from biased AI systems is to promote fairness in algorithmic decision-making processes. This involves ensuring that AI models are trained on diverse and representative datasets, free from discriminatory biases. Additionally, fairness-aware machine learning techniques can be employed to identify and mitigate biases in algorithmic predictions, thereby promoting equitable outcomes for all individuals.

➤ Enhancing Responsibility

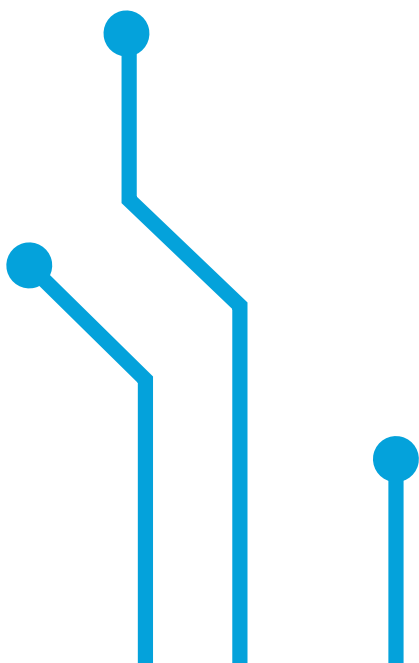
Another important aspect of reducing harm from biased AI systems is to enhance responsibility among developers, organizations, and policymakers. This includes implementing ethical guidelines and best practices for AI development, such as conducting thorough impact assessments to identify potential risks and harms.



Moreover, establishing clear accountability mechanisms and oversight frameworks can help hold individuals and organizations accountable for the ethical implications of their AI deployments. In the next unit of this course, we will explore the concept of Accountability in more detail.

➤ **Encouraging Transparency**

Transparency is essential for making AI systems less harmful by promoting accountability and trust among stakeholders. Transparent documentation of AI algorithms and decision-making processes allows for external scrutiny and validation, ensuring that biases and errors are identified and addressed in a timely manner. Furthermore, fostering open dialogue and collaboration between AI developers, researchers, and affected communities can facilitate greater transparency and understanding of the ethical implications of AI technologies. Unit 4 of this course will dive deeper into the concept of Transparency, as it is one the most fundamental aspects to ensure responsible AI.

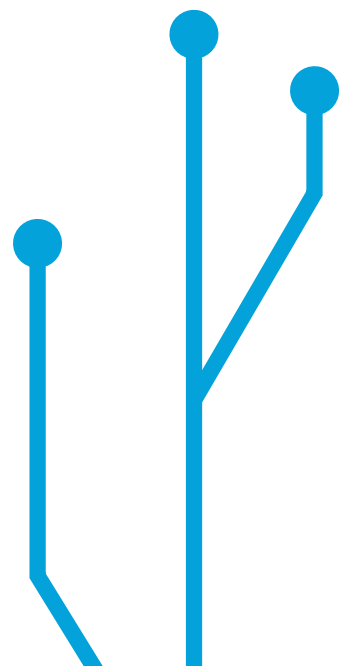


➤ **Ensuring Privacy**

AI systems are powerful tools, but their convenience shouldn't come at the expense of privacy. This strategy focuses on safeguarding your personal information. Developers should collect and use as little data as possible, especially sensitive details. Security measures need to be top-notch to keep information safe. AI systems should also be built to comply with privacy laws and regulations, including the General Data Protection Regulation (GDPR) in Europe, which grants individuals significant control over their personal data.

➤ **Prioritizing Safety**

When it comes to AI, safety should be the top priority. This means putting AI systems through rigorous testing and validation processes before they're released into the real world. The goal is to identify and fix any potential risks or problems that could cause harm. By ensuring AI systems operate reliably and securely, we can safeguard individuals and society as a whole.







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



2022-1-ES01-KA220-HED-000085257