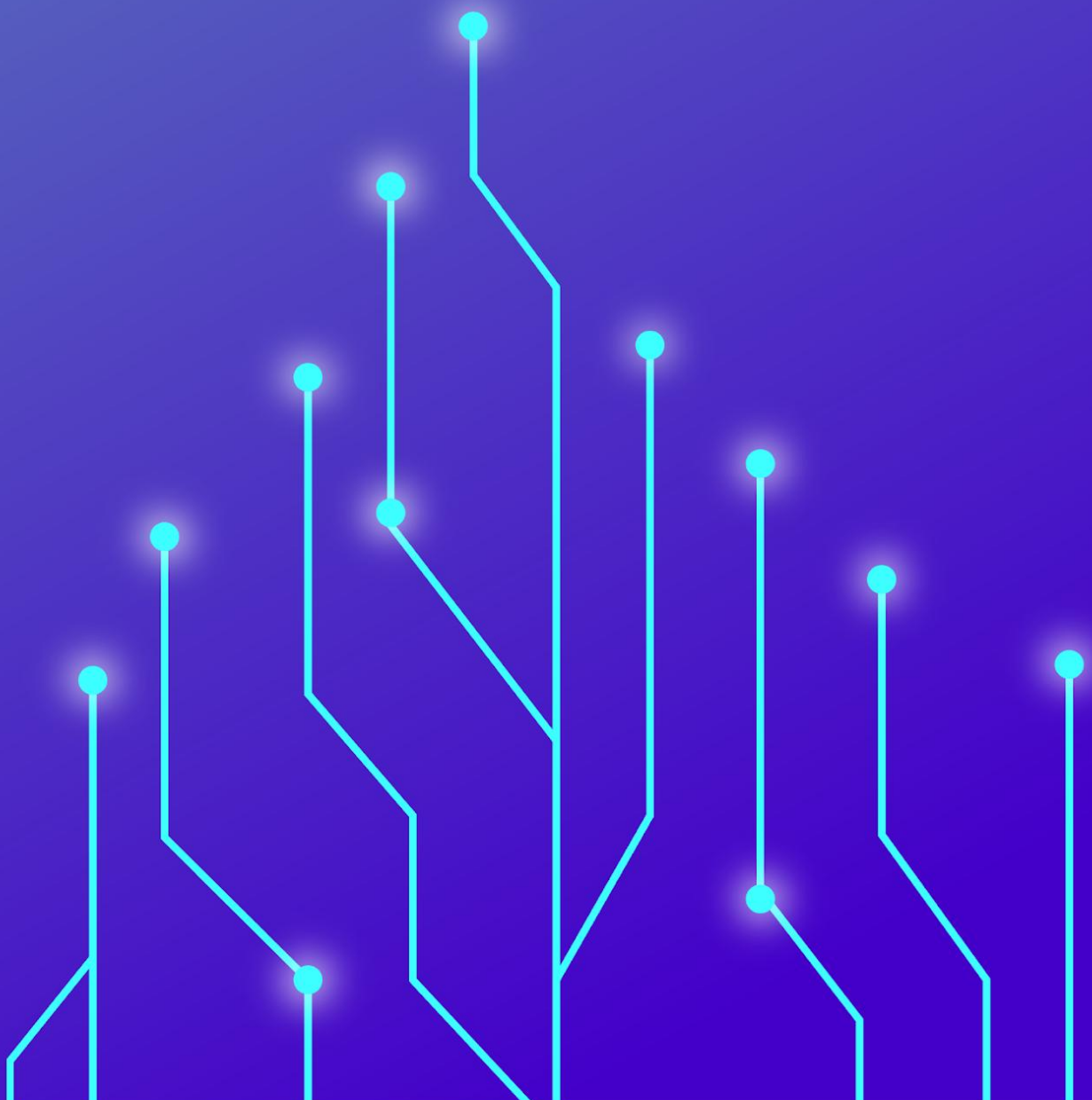
Algorithmic foundations and ethics in AI course: from theory to practice

Handbook

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

Index

CU1 | AI ethics - A practical approach



Index

1. Introduction

TIPS AND RECOMMENDATIONS FOR TEACHERS

This handbook is aimed to guide you through the material for the course unit. The support PPTs contain similar information to what the E-learning covered, but in more detailed form. The PPTs are extensive and are not meant to be covered in their entirety during the interactive sessions. Instead, you can ask at the beginning of the interactive sessions which topics were most difficult to comprehend for the students and use only those support slides during the session. You can ask this by using tools like Mentimeter or Kahoot as an example, adding the course topics from the competence unit content described in bullet points at the end of the introduction section. Furthermore, CU-based case study is aimed to be handled also during the interactive session. Any further suggestions in the handbook are only guidance; feel free to use it as you see fit.

What would happen if Artificial intelligence (AI) would have the power to decide whether you get a loan? Or what happens if AI would judge who will be sentenced for prison? What are the consequences if an autonomous car would have to decide whether to save a pedestrian or the lives of those inside the car? What if an AI-based recruitment system would decide whether to hire a female or male candidate as a flight captain?

The increasing integration of AI into various aspects of society raises significant ethical concerns. From decisions about loans, criminal sentencing, and autonomous vehicles, to hiring practices, the reliance on AI algorithms brings forth questions about ethical principles such as fairness, transparency, and accountability. AI technologies have the potential to perpetuate biases, discrimination, and inequalities if not carefully developed and regulated. Ensuring that AI systems prioritize ethical considerations is essential for promoting justice, equality, and trust in their applications across different domains. (UNESCO, 2023c)

This unit will go deeper into ethical principles, frameworks, and guidelines affecting the development of the AI solutions. AI development lifecycle is examined from problem statement to operation and monitoring of the AI solution. Related principles and stakeholder engagement are considered in each step of the process. In high level, this Course unit handles the following topics.

- AI ethics and its importance in AI solution development
- Ethical principles of AI
- Ethical Frameworks, Guidelines and Toolkits
- High level expert group; Trustworthy AI – Framework
- EU AI Act

- AI Development Lifecycle
- Stakeholder engagement in AI development lifecycle
- AI Development lifecycle with related ethical principles and stakeholders

2. Defining ethical principles, frameworks, and guidelines

Artificial intelligence (AI) is impacting all facets of life, affecting work, leisure, and offering solutions to global issues such as climate change and healthcare access. In the face of AI's widespread integration into economies and societies, the question arises regarding the appropriate policy and institutional frameworks necessary to steer AI's development and application, ensuring societal benefit.

Broad ethical considerations in the field of AI have prompted the creation and publication of numerous approaches to ensure the ethical application of artificial intelligence. (Prem, 2023) Ethical principles, frameworks, and guidelines are commonly used and widely recognized as the core structure in ethical decision-making across many disciplines. They provide a comprehensive approach to navigating ethical issues, from establishing foundational values to guiding specific actions. While often intertwined, they can be described as follows:

Ethical principles provide a moral framework for evaluating and solving ethical dilemmas. Ethical principles form the core values and beliefs that serve as a compass for ethical decision-making and provide a basis for determining what is right and what is wrong in different contexts. (Prem, 2023; Zhou & Chen, 2022)

Ethical frameworks are systematic approaches or models that help individuals and organizations navigate ethical issues. Ethical frameworks are comprehensive models or structured methods that facilitate the examination and resolution of ethical issues. They provide a framework for assessing the ethical consequences of an action and guide the decision-making process, ensuring that ethical considerations are systematically taken into account. (Prem, 2023; Zhou & Chen, 2022)

Ethical guidelines are specific rules, standards and best practices that provide practical guidance on ethical behaviour in various contexts. Ethical guidelines are detailed rules and benchmarks that provide practical advice on how to apply ethical principles in different scenarios. These guidelines help both individuals and organizations to effectively cope with ethical problems and to maintain ethical behaviour in their operations and decision-making. (Prem, 2023; Zhou & Chen, 2022)

3. Common effort to shape the ethical landscape

Ethical principles, frameworks and guidelines come from many different fields and organizations including private companies, research institutions and public sector organizations, representing a common effort to shape the ethical landscape of artificial intelligence. (Jobin et al., 2019; Morley et al., 2020)

Guidance documents and reports on ethical AI are examples of non-legislative policy tools or 'soft law'. In contrast to 'hard law', which includes legal regulations established by legislative bodies to mandate or forbid certain behaviours, ethical guidelines lack legal force but carry persuasive power. These documents are designed to support decision-making processes in specific areas and have been noted to exert considerable influence on practices. (Jobin et al., 2019)

Examples of documents defining the ethical principles of artificial intelligence are:

The OECD AI Principles promote innovative and reliable use of AI that respects human rights and democratic values. Adopted in May 2019, they set standards for artificial intelligence that aim to be practical and flexible. (OECD, 2019)

Beijing Academy of Artificial Intelligence (BAAI) released the "Beijing AI Principles," an outline to guide the research and development, implementation, and governance of AI. ("Beijing Academy of Artificial Intelligence - Beijing AI Principles," 2019; *Beijing Artificial Intelligence Principles*, 2022)

Major industry players such as Google, IBM, Microsoft, and Intel have developed their own ethical guidelines that reflect their unique perspectives and areas of operation. (Jobin et al., 2019; Morley et al., 2020)

Government bodies have not been left behind, with key documents such as the Montreal Declaration and contributions from bodies such as House of Lords Select Committee on Artificial Intelligence and the European Commission Expert Group, each involved in AI ethics regulation and policy. (Morley et al., 2020)

Academic institutions and think tanks, including the Future of Life Institute, IEEE, and AI4People, have added depth to the debate with scientific research and theoretical models. (Floridi et al., 2018; Jobin et al., 2019; Morley et al., 2020)

Combining these efforts reflects a global movement towards responsible AI, and each contribution forms a piece of the larger puzzle of building ethical AI systems that are in line with societal values and norms.

TIPS AND RECOMMENDATIONS FOR TEACHERS

You can ask students to check the above-mentioned sources or some of them, and discuss the differences in the principles they present.

4. Ethical principles

The landscape of ethical principles in AI is rich and diverse, with numerous principles proposed by a variety of actors, including academic institutions, industry leaders, government bodies, and international organizations. The proliferation of these principles reflects a growing consensus on the need for ethical guidance in the development and application of artificial intelligence technologies. (Jobin et al., 2019; Morley et al., 2020; Prem, 2023)

Jobin et al (2019) conducted a comprehensive review of various ethical AI principles and distilled them to identify common themes and elements that recur from different sources. Among the guidelines they reviewed, they found an emerging consensus on key ethical principles for artificial intelligence. Although not universally supported, the convergence principles include openness, fairness and justice, responsibility, reparation, privacy, benevolence, freedom and autonomy, trust, sustainability, human dignity and solidarity.

The common elements identified refer to a set of common concerns and aspirations in the global community. They point towards an emerging consensus on the values that many believe should underpin the development and deployment of artificial intelligence systems. This preliminary agreement forms a fundamental common ground that can serve as a starting point for setting expectations and evaluating results. (Morley et al., 2020)

Morley et al (2020) have acknowledge this collective agreement as a fundamental starting point. They argue that these agreed upon principles characterize an ethically adapted AI as follows:

- Beneficial and respectful to people and the environment (beneficence).
- Robust and secure (non-maleficence).
- Respectful of human values (autonomy).
- Fair (justice).
- Explainable, accountable, and understandable (explicability).

Beneficence

The principle of beneficence is one of the key ethical considerations in the development of artificial intelligence, and it emphasizes the importance of artificial intelligence systems that not only benefit individuals and society, but also contribute to the environment. According to Floridi et al. (2018), the goal is to ensure that artificial intelligence works in a way that promotes general well-being and environmental health.

Jobin et al. (2019) list several key actions to effectively promote artificial intelligence:

- Align AI with human values and ensure that AI systems support and reinforce these values rather than undermine them.
- Involve AI stakeholders in its development process and consider their needs and feedback to improve the impact of AI systems on their lives.
- Actively working to minimize and resolve potential conflicts of interest, using user feedback to guide the development and application of AI.
- Creating new ways to measure human well-being.

TIPS AND RECOMMENDATIONS FOR TEACHERS

Discuss with the students the environmental and societal impacts. For example, TikTok algorithms can provide harmful content if the user shows interest in it, leading to negative mental health effects. Additionally, the significant amount of electricity used to create large language models can leave a considerable carbon dioxide footprint.

Non-maleficence

Ethical principle of non-maleficence is devoted to the prevention of harm. According to Floridi et al. (2018), non-maleficence covers proactive measures against both intentional abuse by humans and unintentional negative actions of AI systems that may affect human behaviour. The core goal is to ensure that AI does not lead to harmful effects regardless of their origin.

To manage risks and safeguard against harm in AI systems, the Jobin et al. (2019) have identified following approaches:

- Using strategic risk management to anticipate and reduce potential risks related to AI systems.
- Implementation of a combination of technical measures and governance policies covering the entire life cycle of AI, aimed at preventing harm before it occurs.

- Certain damages may be unavoidable despite all efforts, which requires a comprehensive risk assessment. This includes measures to reduce and mitigate risks and clearly assign responsibility when harm does occur.

TIPS AND RECOMMENDATIONS FOR TEACHERS

Discuss with students the possible harms caused by AI. Eg. Amazon's AI recruitment tool showed bias against female applicants. Check for more details eg. from here: Insight - Amazon scraps secret AI recruiting tool that showed bias against women | Reuters

Autonomy

The autonomy in the context of AI focuses on maintaining a balance between human decision-making and the influence of AI. Humans need to retain decision-making control and choose when to make their own choices and when to let AI decide. However, humans should always have the ability to take control back from the AI if necessary, meaning they have the final say. (Floridi et al., 2018)

Jobin et al (2019) identify a collection of measures to promote freedom and autonomy in AI systems:

- Improving the transparency and predictability of AI systems so that users can understand and anticipate AI behaviour.
- Improving the general public's understanding of AI technologies.
- Implementation of strong notice and consent protocols to secure individual autonomy when interacting with AI systems.
- Avoiding the collection and distribution of personal data without the express and informed consent of the persons concerned, respecting their privacy and independence.

TIPS AND RECOMMENDATIONS FOR TEACHERS

Discuss with students the possible autonomy issues with AI. Eg. Facebook provides tools for users to exercise control over their content, but can the complexity and opacity of the algorithm sometimes undermine true user autonomy?

Justice

Justice is a multifaceted principle that seeks to utilise AI as a tool to correct past injustices, to ensure that the benefits of AI are distributed fairly, and to protect against the emergence of new harms or societal disruptions. (Floridi et al., 2018)

The realisation of justice in AI, as discussed by Jobin et al. (2019), includes several proactive measures:

- Establishing technical standards and norms.
- Increasing the transparency of rights and regulations.
- Implementation of testing, monitoring and auditing, especially in data protection units.
- Strengthening the rule of law, including appeals and legal remedies.
- Systemic changes are implemented, such as government oversight, versatile teams and inclusive civil society, with a focus on fair distribution of benefits.

TIPS AND RECOMMENDATIONS FOR TEACHERS

Check for example the case with Microsoft's chatbot Tay: In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation - IEEE Spectrum. Referring to justice; the algorithm unfairly prioritized sensational content over meaningful and constructive discourse, leading to an imbalanced representation of views and voices on the platform creating societal disruptions.

Explicability

Explainability is crucial in building and maintaining user trust in AI systems. This means that the processes must be transparent, the capabilities and purpose of artificial intelligence systems must be openly communicated, and the decisions must be explained to those directly and indirectly affected. (High-Level Expert Group on AI (AI HLEG), 2019) Floridi et al (2018) also sees that explicability is central in the application of other ethical principles.

To increase transparency in AI systems, Jobin et al. (2019) list several approaches:

- AI developers and users are encouraged to share more information about their systems, demystifying the processes behind AI decision-making.
- Explanations of AI operations and decisions in ways that are easily understood by non-experts, ensuring that the AI's operations can be audited by humans who are not necessarily experts in the field.
- The use of audits as a way to ensure transparency in AI, which can help identify and address potential bias or ethical issues.

- Establishment of mechanisms that monitor AI performance, engagement with stakeholders and the public to gather diverse perspectives, and mechanisms to support whistleblowing.

TIPS AND RECOMMENDATIONS FOR TEACHERS

Ask the students if they feel informed enough about the decisions made by the AI solutions they are using. For example, Amazon's AI recruiting tool lacked transparency in its decision-making process. Candidates and hiring managers were not provided with clear explanations for why certain candidates were rejected, making it difficult to challenge or address the underlying biases in the system.

NOTE! At the end of the principles section, please utilize the case study provided as a group work.

5. Ethical Frameworks, guidelines, and toolkits

Once the ethical principles have been identified, they should be transformed into actionable frameworks and strategies for guiding AI-based innovations and the practical implementation of AI ethics.

According to Ayling & Chapman (2022), ethical frameworks translate ethical principles into practical measures. This process involves creating guidelines and procedures that developers and adopters of AI technologies can adopt, thereby incorporating ethical considerations into every step from design to real-world applications.

Prem (2023) expands on this by stating that many ethical AI frameworks actively seek to anticipate and address potential ethical issues and risks. By proactively identifying these challenges, frameworks guide developers in creating AI solutions. This proactive approach is essential in a field that is evolving as rapidly as artificial intelligence, where new ethical considerations may emerge as the technology develops and becomes more integrated into different aspects of everyday life.

An ethical framework provides a comprehensive blueprint for establishing regulatory and operational standards that harmonize with ethical norms. As outlined by Floridi & Cowls (2019), such a framework is integral to crafting legislation, formulating rules, setting technical standards, and identifying best practices that are applicable across a diverse array of industries and geographical regions.

Prem (2023, p. 701) defines the essential components of ethical frameworks as follows:

- Basic **concepts** related to the discussion of ethical aspects
- Ethical **principles** (e.g., values)
- **Concern** about how the use and development of artificial intelligence systems adhere to ethical principles
- **Remedies** to address concerns (e.g. strategies, rules and guidelines)

Toolkits and guidelines for applying ethical principles to design, implementation and deployment are very necessary. (Zhou et al., 2020) Ethical toolkits and guidelines are important tools to steer AI innovation in a direction that is in line with ethical norms and values. Ethical toolkits and guidelines are seen as important in bridging the gap between abstract ethical principles and their concrete integration into AI practices.

While the importance of toolkits and guidelines is clear, translating the principles into concrete tools is challenging and many existing frameworks lack application context. (Prem, 2023, p. 702) The reason for the big gap between principles and practise is due to lack of standardization, variability, complexity, and the subjective understanding of the rather abstract principles.

The recent rapid pace of AI development has led to a substantial growth in both the number and diversity of available tools, with new developments emerging constantly. Morley et. al (2019) and Prem (2023) have reviewed hundreds of different tools for addressing the ethical challenges in the AI solutions. Those tools are divided by the AI solution design stages (later defined from chapter H onwards) and by different ethical principles. However, many tools are perceived as difficult to use or are seen to require a substantial amount of effort from their users.

The lists compiled by Morley et. al (2019) and Prem (2023) shows clearly that the availability of the tools and methods is divided unequally between the different principles and design stages. As an example, the tools to measure the beneficence of the solution are numerous at the early stages of the development, while explicability tools are mostly positioned at the later stage of the development process. Check the tools via this link: Applied AI Ethics Typology - Google Docs.

The methods used also vary based on the development stage. Frameworks, libraries, and algorithms often play bigger role in the ex-ante phase, specifically at the design stage of the solution. Some frameworks however reach both ex-ante and ex-post stages (UNESCO, 2023b, p. 5) In the test phase again, audits and metrics are predominantly presented, while in the post ante face mainly declarations, labels and licences are used as approaches. (Prem, 2023, p. 709)

The study also revealed that only few tools addressed the solution's impact on an individual or on a society as whole. This might be due to the reason that it is challenging to change the complex human behaviour into simple and general

design tools. However, it would be essential to address the human and society impact to achieve an acceptance and social preferability of the AI solutions. (Morley et al., 2020, p. 2156)

6. From principles to practise; Trustworthy AI framework by HLEG

Although there is consensus that AI must be ethical, debates continue about the exact definition of "Ethical AI" and the ethical obligations and technical criteria needed to implement it. (Leijnen et al., 2020)

The European Commission has rolled out a comprehensive strategy for the advancement of artificial intelligence (AI) within Europe, placing a strong emphasis on ensuring that the AI developed and used is not only advanced from a technological standpoint but also adheres to ethical standards and is secure.

To effectively implement this strategy and its underlying principles, the European Commission established the High-Level Expert Group on Artificial Intelligence (AI HLEG). This group was tasked with the development of AI Ethics Guidelines and Policy and Investment Recommendations, which are meant to guide the ethical development, deployment, and use of AI across Europe, fostering innovation while safeguarding fundamental values and rights. [High-Level Expert Group on AI (AI HLEG), 2019]

Trustworthy AI framework is designed to ensure that AI technologies are developed and deployed in a way that is ethically sound, respecting fundamental rights and ensuring robustness and reliability. The framework for Trustworthy AI is composed of three components, each designed to guide the development and deployment of AI in an ethical and reliable manner. [High-Level Expert Group on AI (AI HLEG), 2019]

Foundations of Trustworthy AI

This segment establishes the core principles underlying Trustworthy AI, embracing a fundamental rights based approach. It delineates the ethical principles essential for ensuring that AI systems are developed and utilized in a manner that is both ethical and robust.

Ethical principles form the bedrock upon which Trustworthy AI must be built, ensuring adherence to ethical standards. These ethical principles are (i) Respect for human autonomy (ii) Prevention of harm (iii) Fairness and (iv) Explicability.

Realizing Trustworthy AI

Building on the ethical principles outlined, the second part of the framework translates these principles into seven key requirements that AI systems are expected to embody and uphold throughout their life cycle. This transformation of ethical principles into practical requirements ensures that the foundational

ethics are not just theoretical ideals but are actively integrated into every stage of an AI system's development and operation.

The list of requirements include: Human agency and oversight: fundamental rights, human agency and human oversight; Technical robustness and safety: resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility; Privacy and data governance: respect for privacy, quality and integrity of data, and access to data; Transparency: traceability, explainability and communication; Diversity, non-discrimination and fairness: the avoidance of unfair bias, accessibility and universal design, and stakeholder participation; Societal and environmental wellbeing: sustainability and environmental friendliness, social impact, society and democracy; Accountability: auditability, minimisation and reporting of negative impact, trade-offs and redress.

Assessing Trustworthy AI

The third part of the framework provides a concrete and comprehensive list for the assessment of Trustworthy AI, designed to operationalize the aforementioned requirements.

Known as the “Assessment List for Trustworthy AI (ALTAI)”, this tool provides a comprehensive and dynamic checklist that serves as a roadmap for developers and deployers to integrate these principles into their AI applications. ALTAI facilitates this process by offering a set of concrete steps for self-assessment, thereby ensuring that AI technologies are developed in a manner that maximizes user benefits while minimizing exposure to unnecessary risks. (High-Level Expert Group on AI (AI HLEG), 2020)

This non-exhaustive assessment list serves as a practical guide for AI practitioners, offering specific guidance on how to implement these requirements effectively. Importantly, this assessment is meant to be adaptable, allowing for customization according to the specific application of each AI system, thereby ensuring relevance and efficacy in diverse contexts.

7. From principles to law: EU AI ACT

The AI Act regulates the application of artificial intelligence within the EU, marking the first extensive AI legislation globally. (European Parliament, 2023) The purpose of AI ACT is to improve the functioning of the internal market and promoting the uptake of human centric and trustworthy AI, while ensuring a high level of protection of health, safety, fundamental rights, including democracy, rule of law and environmental protection against harmful effects of artificial intelligence systems in the Union and supporting innovation. (Future of Life Institute, 2024)

EU Commission has organized AI regulations within a risk-based framework, creating a system of risk categories. AI applications will be regulated according to the level of risk they present. (European Commission, 2023). The risk-based approach outlined in the EU AI Act categorizes AI systems into four levels based on their potential impact: unacceptable risk, high risk, limited risk, and low or minimal risk. (European Commission, 2023; Hillard & Gulley, 2023)

Minimal Risk

The majority of AI systems fall into this category, presenting minimal or no risk to citizens' rights or safety. Examples include AI-enabled recommender systems and spam filters. Companies operating such systems are not obligated to comply with stringent requirements but may voluntarily commit to additional codes of conduct to enhance transparency and accountability.

Limited Risk

This category pertains to AI systems pose only limited risk to users' rights and safety. Examples include chatbots and deep fakes. While not subject to strict regulatory requirements, providers are mandated to ensure users are aware when they are interacting with AI systems and label AI-generated content accordingly. This ensures transparency and awareness regarding the use of AI technologies.

High-Risk

AI systems identified as high-risk must adhere to strict requirements to mitigate potential risks. These requirements include robust risk-mitigation systems, high-quality data sets, activity logging, detailed documentation, clear user information, human oversight, and strong cybersecurity measures. Regulatory sandboxes are established to foster responsible innovation and facilitate the development of compliant AI systems. Examples of high-risk AI systems include critical infrastructures, medical devices, educational institution access systems, and certain applications in law enforcement and border control.

Unacceptable Risk

AI systems posing a clear threat to fundamental rights are prohibited. This includes systems that manipulate human behaviour to undermine free will, such as toys encouraging dangerous behaviour in minors, or those enabling government or corporate 'social scoring.' Predictive policing applications and certain uses of biometric systems, such as emotion recognition in workplaces, are also prohibited.

TIPS AND RECOMMENDATIONS FOR TEACHERS

1) Discuss with the students if they feel that the EU law will be stricter than other laws and what will it cause. Food for thoughts for example from this article: https://news.bloomberglaw.com/artificial-intelligence/eu-embraces-new-ai-rules-despite-doubts-it-got-the-right-balance
2) If thinking about the recent discussion around TikTok and how harmful it is, in which risk category would it fall?

8. AI Development Lifecycle

As explored in Chapter E, specific ethical principles become more relevant at certain stages of AI system development, with various tools being utilized accordingly. Understanding the development process from a broad perspective is crucial. (Prem, 2023, p. 703) In this course, we adopt the term 'AI Development Lifecycle' to encapsulate this comprehensive view of the development stages.

The fast-paced industry of AI solutions sometimes witnesses project failures due to inadequate project management processes. A well-defined AI lifecycle, considering all developmental steps, can enhance the success rate of AI solutions. While there are different versions of AI lifecycle descriptions, our approach synthesizes the models proposed by Morley et al. (2020), Prem (2023), and the CRISP model by the Data Science Process Alliance (Hotz, 2018) as presented by Rochel & Evequoz (2021).

The AI Development Lifecycle describes the journey an AI system takes from initial conception to final deployment and maintenance. Each phase is vital to ensuring a successful outcome. The design lifecycle is iterative, meaning revisiting previous stages may be necessary if certain variables do not perform as expected. Although ethical considerations may vary across different stages, integrating these considerations into every phase of the process is essential. (Saltz, 2023)

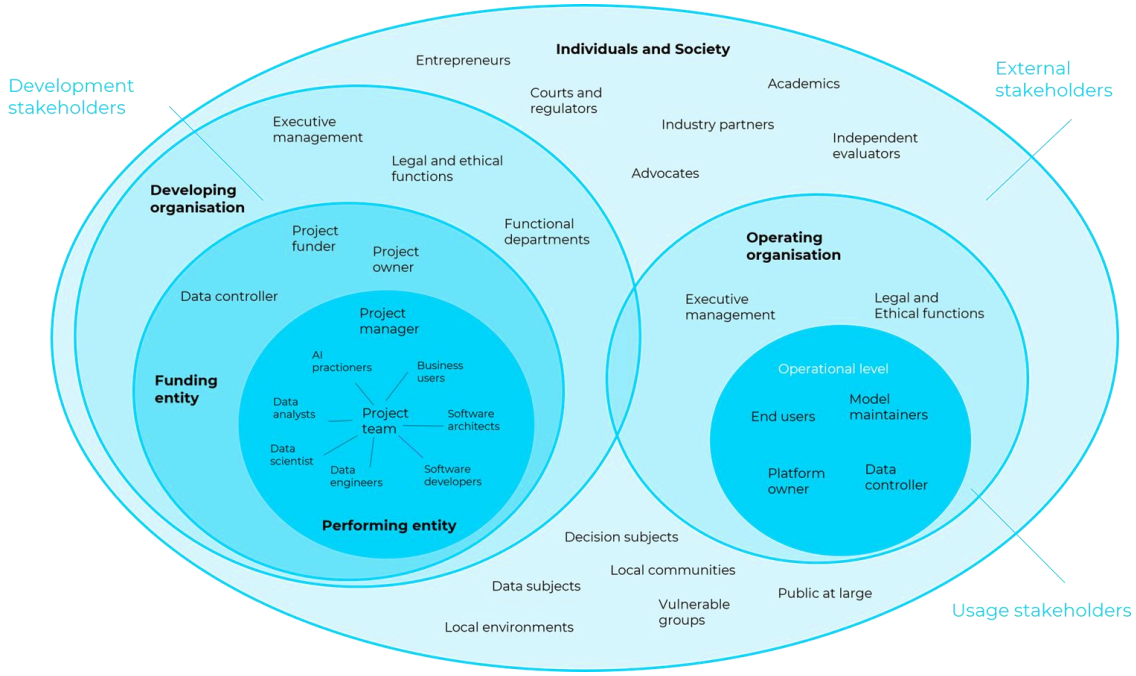
Picture: Authors of the document. Modified as a combination of the different models described by the Data Science Process Alliance (2023), Morley et.al' (2019), Prehm's (2022) and Rochel et. all (2020) articles.

9. Stakeholder engagement in AI Development lifecycle

Involving a diverse range of stakeholders in the AI development process is crucial to thoroughly evaluate the ethical impacts of an AI project from multiple perspectives. Consulting stakeholders, including those directly impacted by the AI solution, is essential for building trustworthy systems (Morley et al., 2020) and for a comprehensive assessment of the project's impacts (UNESCO, 2023a, p. 15). This chapter provides an overview of the different stakeholders involved in AI development projects and describes their roles at various stages of the AI development lifecycle.

The stakeholders are divided into development stakeholders, usage stakeholders, and external stakeholders. The development and usage stakeholders are actively involved stakeholders that are for instance the decision-makers who have the power to decide, and the designers with the expertise. The external stakeholders are not involved in the AI project but may affect or be affected by it or be consulted. (Miller, 2022)

Next, the key roles of various stakeholders in an artificial intelligence development are examined and classified into three categories: development, use, and external stakeholders. Also, the specific responsibilities and contributions of each group at different stages of the AI development lifecycle are explored. Understanding these roles is essential for navigating the complexities of AI development and ensuring the success of the project.



Picture: G.J. Miller (2022)

Development stakeholders

In the realm of AI project development, Miller (2022) divide stakeholders into three principal categories: the organization, the funding entity, and the performing entity, each playing a distinct role in the project's lifecycle. The organization benefits from the outcomes of the AI project, leveraging the advancements and innovations for its growth and success. The funding entity, crucial for financial support, not only funds but also governs the project, ensuring that investments align with project goals and governance standards. The performing entity is pivotal, taking charge of generating the project's outputs and crafting the AI system, embodying the project's technical execution. (Miller, 2022)

Within the organization, the roles span from the developing organization, which sponsors and defines the project's scope, to executive management, which

oversees the project's trajectory. Functional departments provide essential support, whereas legal and ethical functions ensure the project aligns with regulatory standards and ethical norms. The project management group plays a supportive role, assisting project managers with administrative tasks to streamline project execution (Zwikael & Meredith, 2018).

The funding entity encompasses roles such as the project funder, who supplies financial resources, and the project owner, who imparts strategic direction and potentially chairs the steering committee, ensuring project alignment with strategic goals. In scenarios where the funder is not directly involved, accountability is delegated to ensure continuous oversight. The data controller, a role often associated with the funding entity, manages data usage, addressing compliance with stringent regulatory and legal frameworks to avoid potential penalties (Miller, 2022).

At the heart of the performing entity are the project managers and the project team, responsible for delivering the project's outputs as per the approved plan. This collaboration ensures that the project advances from conception to fruition, with each team member contributing to the development and successful deployment of the AI system (Zwikael & Meredith, 2018).

This structured organization of stakeholders and their roles underscores the collaborative and multifaceted approach necessary for the successful development and implementation of AI projects, highlighting the importance of clear roles, responsibilities, and compliance with legal and ethical standards.

Usage stakeholders

In the context of AI projects, usage stakeholders play a crucial role, particularly at the organizational level where the operating organization functions as a client stakeholder that acquires the AI solution. Similar to the development organization, the operating organization encompasses roles such as executive management and legal and ethical functions, highlighting the parallel structures in both development and operational phases (Miller, 2022).

At the operational level, the usage stage involves a variety of key players including data custodians, end users, platform owners, and model maintainers. End users, whether they are individuals or groups, interact with the AI system either through work or service contracts with the operating organization or as consumers, directly engaging with the system's functionalities (Miller, 2022).

Model maintainers undertake the critical task of continuous management, updating, and enhancing the machine learning models that drive the AI system, ensuring its efficiency and effectiveness. Data controllers are charged with defining the objectives and methods for processing personal data within the system, ensuring compliance with data protection laws. Lastly, platform owners bear the responsibility for the AI system's overall infrastructure and its

deployment, maintaining the foundational aspects on which the AI operates (Miller, 2022).

External stakeholders

Miller (2022) categorizes External stakeholders in AI projects into three main sub-groups: individuals, society, and representatives, each with distinct roles and contributions to the project.

The individuals sub-group encompasses data subjects, decision subjects, and workers who interact with the development or operating organization through agreements like service or employment contracts. These stakeholders include individuals directly involved in or impacted by the project. These stakeholders play crucial roles as data subjects, decision subjects, and workers within the project (Miller, 2022).

The society sub-group captures the broader impact on the general populace, including local communities, the environment, and vulnerable groups such as people with disabilities, minority groups, minors, and the public at large. This segment is directly influenced by the development and utilization of the AI system, emphasizing the societal reach of AI projects (Miller, 2022).

Representatives, forming the third sub-group, are distinguished from formal representatives. While representatives include groups and organizations like entrepreneurs and independent evaluators (reporters, auditors, or reviewers) who may not be directly affected by the AI system but play a role in its development or evaluation, formal representatives such as courts, laws, regulators, and labour unions represent legal and regulatory oversight and governance within the project (Miller, 2022; Rodrigues, 2020).

Stakeholder engagement in AI solution development

Co-creation refers to collaborative engagement with stakeholders, including end users or customers, throughout the design process. This approach facilitates the exchange of insights among participants with varied roles and expertise. (Interaction Design Foundation - IxDF, 2021; Sanin, 2020)

The practice of co-creation can be executed via facilitated workshops that encompass a variety of activities, such as ice-breakers, role-playing, journey mapping, brainstorming, and prototyping. By embracing co-creation, designers are enabled to deeply understand the needs and desires of customers or end users, thereby crafting solutions that are considerate of diverse user perspectives. This inclusive method ensures that the developed services are well-aligned with the actual requirements of those they aim to serve. (Interaction Design Foundation - IxDF, 2021; Sanin, 2020)

10. AI Development Lifecycle: Problem definition and business understanding

TIPS AND RECOMMENDATIONS FOR TEACHERS

This section will explain in detail the development lifecycle of an AI solution, its ethical implications and related stakeholders. To foster a vivid discussion, please use an example case to go through the whole lifecycle. As an example, you can use the AI based recommendation system on what to purchase via this link: https://www.datascience-pm.com/ai-lifecycle/#An_Example_Ai_Project_Life_Cycle

Like any projects, the AI lifecycle should also start with the problem definition. Well defined problem statement gives a deep understanding of customer's need and thereby a direction for the project. At this point, a feasibility study should be conducted to understand whether AI is the most appropriate solution to address the problem. Also, research and stakeholder engagement are necessary to understand the problem thoroughly. (De Silva & Alahakoon, 2022, pp. 4–5)

Ethical principles in the Problem definition and business understanding stage

In the initial phase of a project, it's crucial to consider principle ethical aspects, as they lay the groundwork for the project's path and subsequent societal impact. Particularly at this stage, the principles of beneficence and non-maleficence are considered paramount. (Prem, 2023, p. 703)

Regarding beneficence, the focus is on assessing whether the AI system actively fosters individual well-being and societal benefit. The system's objectives must be transparent, aiming for tangible benefits while upholding fundamental rights and considering environmental sustainability. It is essential to engage a variety of stakeholders, ensuring that the perspectives of those affected by the AI solution are heard and integrated. (De Silva & Alahakoon, 2022, p. 5; Morley et al., 2020, p. 2151)

Conversely, non-maleficence involves ensuring that the AI system does not inflict harm on individuals or communities. This requires a proactive approach to identifying and mitigating potential adverse effects, vigilantly safeguarding the physical and mental well-being of people impacted by the system. (De Silva & Alahakoon, 2022, p. 5; Morley et al., 2020, p. 2151) General safety concerns must be proactively addressed and built into the system's design (UNESCO, 2023b, p. 18).

From the standpoint of justice, it's imperative to scrutinize the AI system's broader implications on institutions, democracy, and societal structures. It is vital to consider whether the project could inadvertently create biases, favouring certain groups over others, or encroach upon individual autonomy. Each of these aspects demands careful consideration to ensure the AI system contributes to a fair and equitable society. (Morley et al., 2020, p. 2151)

Stakeholders in the Problem definition and business understanding stage

At this stage, it is essential to involve at least the end-users, who are the primary users of the AI solution. Additionally, AI developers should be included early on to gain a thorough understanding of the problem they will be tasked with solving. Ethicists or social scientists should also be involved to assess the potential human and societal impacts of the solutions. Domain experts are also crucial to include to increase the knowledge around the industry where AI is being applied to. (De Silva & Alahakoon, 2022)

11. AI Development Lifecycle: Design stage

During the design stage, the foundational work done in the problem definition stage evolves into a comprehensive plan for deployment of the AI system project. This phase includes the definition of project goals and desired outcomes, the development of a project plan with timelines and budget considerations, and an understanding of stakeholder requirements. (De Silva & Alahakoon, 2022, p. 4)

Key activities include crafting a requirements specification that formulates the AI system's functions and the metrics for project success. Additionally, strategies for data mining are refined, including for example the decisions regarding the use of pre-trained models. Throughout this stage, ethical considerations are integral and must be considered in every aspect of the system's design. (De Silva & Alahakoon, 2022, p. 3)

Ethical principles in the Design stage

During the AI solution's design phase, it is crucial to employ proactive "what if" strategies to foresee and mitigate potential challenges throughout the AI solution's lifecycle. By posing critical inquiries such as, "What if the data used is biased?" and exploring necessary mechanisms to detect and address unjust behaviour, developers can preventively tackle these issues. The establishment of concrete steps is imperative to equip developers with the knowledge to avert biases related to e.g. race, colour, descent, gender, age, language, religion, political opinion, national or ethnic origin, social background, economic status, birth conditions, or disability. Furthermore, it is essential to set up robust

communication channels, reporting protocols, and measures that are swift to react to biases and the adjustments they might require. (Morley et al., 2020, pp. 2151, 2155; UNESCO, 2023b, pp. 13, 17)

In addition to these precautions, the design requirements should clearly articulate methods to enhance the transparency of the process, thereby improving explicability. This approach underscores the significance of making the AI system's operations understandable and accountable (Morley et al., 2020, pp. 2151, 2155; UNESCO, 2023b, pp. 13, 17).

Furthermore, Prem (2023, p. 703) emphasizes the value of involving stakeholders during this stage and the critical role of human oversight mechanisms. This inclusion ensures that a diverse range of perspectives contributes to the AI's development, promoting a more equitable and informed design process.

Stakeholders in the Design stage

In this phase it is important to involve software architects as they play a crucial role in laying the technical foundation of the AI solution and consider both current requirements and future scalability (Peter, 2024). UX designers are essential for creating interfaces that facilitate user adoption and satisfaction and ensuring that users can interact seamlessly with the AI system (White, 2023). AI practitioners need to collaborate closely with UX designers to integrate AI components seamlessly into the user interface. Ethics experts address ethical concerns in AI development. Legal and ethical functions are needed to reduce the legal risks associated with AI projects.

According to De Silva and Alahakoon (2022), it is recommended to use a participatory design approach that involves people and communities affected by the AI model and its decisions, fostering transparency, inclusivity, and alignment with societal values throughout the project lifecycle.

12. AI Development Lifecycle: Data collection, understanding and preparation

After setting the objective and devising a strategy, the next phase involves acquiring and preparing relevant data. This includes:

- Gathering the data
- Defining and categorizing it
- Conducting exploratory analysis
- Validating its relevance
- Selecting essential information
- Cleaning, reconstructing, integrating, and formatting it to fit the project's goals

Addressing and compensating for missing values is crucial. Understanding the assumptions behind the existing data and ensuring clarity in how datasets are used in decision-making processes is essential. Maintaining data quality is paramount, especially when the data is incomplete or ambiguous. (Rochel & Evéquoz, 2021, pp. 614–617)

While this phase is notably time-intensive, it is also foundational. The reliability and performance of the resulting AI solution are directly correlated to the quality and precision of the underlying data. (Saltz, 2023)

Ethical principles in the data collection, understanding and preparation stage

The data collection phase comes with inherent risks, particularly concerning the principle of non-maleficence. Ensuring data privacy and confidentiality is critical during this phase. AI systems must uphold data privacy consistently, from inception through the entire lifecycle of the solution. Legal regulations, such as the Global Data Protection Law, typically govern privacy matters. (De Silva & Alahakoon, 2022, p. 5)

Biases within the data can lead to the overrepresentation of certain demographics, so it's essential to scrutinize and assure the quality and integrity of the data. Additionally, datasets are exposed to cyber threats, necessitating robust security measures to shield them from potential attacks. (UNESCO, 2023a, pp. 9, 18)

Being transparent about the data collection processes, including what data is being collected, for what purposes, and how it will be used in the AI system is also valid at this point. (De Silva & Alahakoon, 2022, p. 6)

Stakeholders in the data collection, understanding and preparation stage

Involving data scientists during this phase is crucial, as they play a pivotal role in discerning patterns and trends within the data, transforming it into actionable knowledge (Miller, 2022). Similarly, AI engineers are indispensable for assessing and verifying the data's integrity (Rochel & Evéquoz, 2021), while data analysts enhance the project by comprehending the data's characteristics and advising on effective data preparation strategies. Additionally, data engineers' contribution is vital for the preparation and maintenance of the data, ensuring its quality (Miller, 2022).

The inclusion of a Data Protection Officer (DPO) is important to ensure compliance with legal and ethical standards concerning personal data (European Data Protection Supervisor, 2024).

Ethicists play a significant role by integrating ethical considerations throughout the data handling and selection process. Engaging these diverse stakeholders

equips the AI project with ethically sourced and meticulously prepared data, establishing a foundation of trust and reliability for the AI solution.

13. AI Development Lifecycle; Model development and training

During the model development and training phase, the creation of an AI model is specifically aimed at tackling the pre-identified challenge. This phase includes the careful selection, setup, and intended use of algorithmic tools, making it essential to thoroughly document and assess the implications of these choices, particularly identifying any potential drawbacks or necessary compromises. The iterative nature of this stage is a defining characteristic, involving multiple rounds of refinement to improve the model's accuracy and effectiveness.

The process of selecting algorithmic tools often involves navigating between conflicting objectives. A prime example is the challenge of filtering spam emails, where there's a need to balance the goal of minimizing spam in users' inboxes against the risk of incorrectly categorizing legitimate emails as spam. Achieving the optimal balance is critical to enhancing the model's efficacy without compromising its reliability or user trust. (Rochel & Evéquoz, 2021, pp. 617–618; Saltz, 2023)

Ethical Principles in the model development and training stage

While technical details often take precedence, this stage offers a good opportunity for ensuring explicability and interpretability of the chosen model, as well as for assessing whether any biases exist. (Prem, 2023, p. 703) The development process must be clear and comprehensible to all involved parties, necessitating thorough documentation of data sources, model choices, and the logic that informed these decisions. (Morley et al., 2020, p. 2151)

When necessary, trade-offs arise, particularly as efficiency often involves compromises, a systematic approach to recognize and openly assess their effects is essential. AI engineers must be prepared to rationalize their choices, compelling them to account for and mitigate any adverse consequences on those impacted. (Morley et al., 2020, p. 2151)

Stakeholders in the model development and training stage

AI or machine learning scientists play a pivotal role in transforming problem definitions into initial AI model prototypes (De Silva & Alahakoon, 2022), marking their critical involvement as the project advances into the model development and training phase. During this phase, AI engineers become indispensable due to their ability to identify and employ the most suitable algorithmic tools that are

specifically aligned with the project's objectives (Rochel & Evéquoz, 2021). Furthermore, data scientists are at the core of this phase, steering the development and continuous refinement of machine learning models towards greater accuracy and efficiency.

The inclusion of ethics experts is paramount to ensure the embedding of ethical considerations within the development process, aiming to avert any potential adverse impacts. Additionally, the insights provided by business users are invaluable, guaranteeing that the evolving models are in concordance with business strategies and meet the expected outcomes, thereby bridging the gap between technical innovation and practical business applications.

14. AI Development Lifecycle: Model evaluation and testing

After the AI model has been developed and trained, its effectiveness needs to be tested and evaluated against predefined goals and success criteria. This evaluation not only measures the model's performance but also examines the underlying assumptions and selection criteria of the datasets used. In instances where the model does not meet expectations, adjustments may be required, whether in the model's parameters, its overall architecture, or the datasets it relies on. Transparent communication of any shortcomings with the project leader and all relevant stakeholders is crucial, given the potential significant impact on individuals involved. Based on the evaluation outcomes, the team must decide on the subsequent actions, which could involve further iterations for refinement or proceeding to the deployment phase, ensuring that every step is clearly aligned and agreed upon. (Hotz, 2018; Rochel & Evéquoz, 2021, p. 619)

Ethical principles in the Model evaluation and testing stage

At this phase, testing the model for its adherence to ethical guidelines is paramount, ensuring it aligns with the critical principles established at the beginning. The development and refinement of audits and metrics for auditability stand as essential tasks. It's important to incorporate key ethical principles such as justice, beneficence, and non-maleficence. Assessing the model's resilience to potential security threats and setting up a strategy for unexpected challenges are crucial steps. There should be clear mechanisms for corrective action in case of adverse outcomes. A thorough evaluation encompassing privacy, social impact, potential biases, and more is necessary, with a commitment to minimizing and transparently reporting any negative effects. (De Silva & Alahakoon, 2022, p. 8; Morley et al., 2020, p. 2151; Prem, 2023, p. 703)

Transparency is essential, not only concerning the technical workings of the model but also regarding the justifications for human decisions within the project, ensuring explicability (De Silva & Alahakoon, 2022, p. 8; Morley et al., 2020,

p. 2151). Additionally, it's crucial to acknowledge that engineers might face pressure to accelerate the development process for deployment, potentially introducing further risks (Rochel & Evéquoz, 2021, p. 618).

Stakeholders in the model evaluation and testing stage

In this phase, AI engineers play a vital role by interpreting the results from modelling, offering essential insights into the outcomes (Rochel & Evéquoz, 2021).

The Quality Assurance (QA) Teams are crucial for confirming that the models adhere to quality standards and are devoid of errors or unintended results. Involving ethics experts is key to guaranteeing the ethical deployment of AI models, ensuring they are used responsibly. User Advocates or Representatives contribute significantly by aligning the AI models with user expectations and requirements, thereby increasing user satisfaction and acceptance.

Additionally, the Legal and Compliance Teams are indispensable in addressing legal risks, making sure that the AI models are in compliance with all applicable laws and regulations. (Moore, 2023)

15. AI Development Lifecycle: Deployment

After its successful development, the AI solution is deployed in a production environment aimed at solving real-world problems. The scope and method of implementation depend on the solution's specific targets and the systems or applications it is designed for. Typically, the rollout begins with a pilot phase involving a select group of experts and early adopters who can provide initial feedback (De Silva & Alahakoon, 2022, p. 8).

The AI solution might be integrated with existing systems, serve as the foundation for a new application or service, or its findings could be shared through offline methods like managerial reports. It's crucial to establish a feedback mechanism to monitor its performance and impact, facilitating continuous improvements based on real-world usage (Saltz, 2023).

To address potential risks, comprehensive risk management strategies are implemented. Operational and monitoring measures are put in place to ensure the AI solution operates efficiently after deployment. A thorough project review is conducted at the end, resulting in a detailed report summarizing the project's outcomes and lessons learned (De Silva & Alahakoon, 2022, pp. 8–9).

Ethical Principles in the Deployment stage

Morley (2020) highlights that ethical considerations often receive insufficient attention during the deployment phase of the design process. Distilling complex

human behaviours into tools that are both understandable and transparent presents significant challenges. Yet, creating such tools is crucial for enhancing trust in AI systems. Therefore, explicability emerges as a key ethical principle at this stage.

Furthermore, ensuring autonomy is essential by informing end-users when they are interacting with an AI-driven process. This transparency is crucial to avoid over-reliance on AI systems. UNESCO (2023b, p. 36) underscores the need for effective mechanisms in four key areas: system awareness, robust audit procedures, clarity in algorithmic reasoning, and measures to gauge impact, like channels for appeals or complaints, which should include protections for whistle-blowers. Additionally, the critical role of human oversight cannot be overstated, with the necessity for provisions that allow human intervention and supervision of AI systems to ensure their ethical deployment (Morley et al., 2020, p. 2151).

Stakeholders in the deployment stage

AI or machine learning engineers are pivotal in transitioning the prototype AI model into a fully deployed service or solution accessible to stakeholders and end-users (De Silva & Alahakoon, 2022).

Cybersecurity Experts conduct a risk assessment to secure the AI system (Junklewitz et al., 2023, pp. 9–12). End-users contribute significantly by highlighting any issues or shortcomings that might not have been evident during controlled environment testing. Additionally, User Advocates or Representatives are instrumental in gathering feedback regarding the deployed AI solution's usability and performance. Their role ensures that the solution aligns with user expectations and requirements, facilitating improvements where necessary.

16. AI Development Lifecycle: Operation and monitoring

Ongoing maintenance, updates, and evaluations are essential for the longevity and effectiveness of an operational AI system. Constant performance reviews are conducted to ensure the system meets expected standards, while continuous monitoring of user interactions offers valuable insights into real-world application and potential areas for enhancement. Regular updates with new data are critical to keep the model relevant and accurate. User feedback is instrumental in the continuous refinement of the AI model, highlighting the importance of iterative improvements as part of the maintenance cycle, symbolizing progress rather than failure.

Additionally, measuring the system's return on investment (ROI) and other critical metrics is vital to assess its success in achieving predefined objectives. (De Silva & Alahakoon, 2022, p. 9; Saltz, 2023)

Ethical principles in the operation and monitoring stage

In this phase, performance monitoring extends beyond technical metrics to include compliance with ethical standards. Regular reassessment of all principles is critical, especially considering possible changes to datasets and model structure. Stakeholder engagement is key even post-deployment, with their input essential for ongoing enhancements. Feedback should be actively sought from users, organizations, and the broader society. Structured mechanisms should be in place to consistently gather and incorporate this feedback into continuous improvement efforts. (High-Level Expert Group on AI (AI HLEG), 2019, p. 19; UNESCO, 2023b, p. 15)

Stakeholders in the operation and monitoring stage

During the operation and monitoring phase, DevOps and IT Operations Teams are crucial for monitoring system performance and maintaining continuous availability (Immersant Data Solutions, 2023). Independent evaluators, such as, safety certifiers, can assess the security of the developed AI solution (Miller, 2022). Additionally, expert panels, steering committees, or regulatory bodies conduct reviews of the project, covering both technical and ethical considerations (De Silva & Alahakoon, 2022).

17. Conclusion

This unit has delved into the ethical principles, frameworks, and guidelines that are pivotal in shaping the development of AI solutions. It explored the AI development lifecycle, from the initial problem statement through to the operation and monitoring of the AI solution, with a keen focus on the integration of related principles and stakeholder engagement at every stage.

The ethical AI development highlights the importance of incorporating ethical considerations throughout the entire AI development lifecycle. This approach ensures that AI technologies not only contribute positively to society but also address potential risks such as biases and privacy concerns effectively. Moreover, the significance of engaging a broad spectrum of stakeholders was emphasized, illustrating that successful AI development requires input beyond technical expertise to include diverse perspectives. This inclusive approach enriches the development process, ensuring AI solutions are aligned with societal values and achieve wider acceptance.

Furthermore, the need for ongoing monitoring and iterative refinement of AI systems post-deployment was underscored as essential for maintaining their ethical integrity and ensuring their sustained success. By adopting a proactive stance towards innovation, developers can identify and address ethical challenges as they arise, thereby navigating the complexities of AI development with a commitment to ethical standards and responsible innovation.

18. References

Ayling, J., & Chapman, A. (2022). Putting AI ethics to work: Are the tools fit for purpose? AI and Ethics, 2(3), 405–429. https://doi.org/10.1007/s43681-021-00084-x

Beijing Academy of Artificial Intelligence—Beijing AI Principles. (2019). Datenschutz Und Datensicherheit, 10.

Beijing Artificial Intelligence Principles. (2022, January 10). International Research Center for AI Ethics and Governance. https://ai-ethics-and-governance.institute/beijing-artificial-intelligence-principles/

De Silva, D., & Alahakoon, D. (2022). An artificial intelligence life cycle: From conception to production. Patterns, 3(6), 100489. https://doi.org/10.1016/j.patter.2022.100489

European Commission. (2023, September 12). Commission welcomes political agreement on AI Act [Text]. European Commission - European Commission. https://ec.europa.eu/commission/presscorner/detail/en/ip_23_6473

European Data Protection Supervisor. (2024, April 9). Data Protection Officer (DPO) | European Data Protection Supervisor. https://www.edps.europa.eu/data-protection/data-protection/reference-library/data-protection-officer-dpo_en

European Parliament. (2023, June 8). EU AI Act: First regulation on artificial intelligence. Topics | European Parliament. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review. https://doi.org/10.1162/99608f92.8cd550d1

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds and Machines, 28(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Future of Life Institute. (2024). The AI Act Explorer | EU Artificial Intelligence Act. https://artificialintelligenceact.eu/ai-act-explorer/

High-Level Expert Group on AI (AI HLEG). (2019). Ethical Guidelines of Trustworthy AI. High-Level Expert Group on Artificial Intelligence. European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

High-Level Expert Group on AI (AI HLEG). (2020). Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment. European Commission.

Hillard, A., & Gulley, A. (2023, December 9). What is the EU AI Act? https://www.holisticai.com/blog/eu-ai-act

Hotz, N. (2018, September 10). What is CRISP DM? Data Science Process Alliance. https://www.datascience-pm.com/crisp-dm-2/

Immersant Data Solutions. (2023, June 9). DevOps: Bridging the Gap between Development and Operations | LinkedIn. https://www.linkedin.com/pulse/devops-bridging-gap-between-development-operations/

Interaction Design Foundation - IxDF. (2021, October 27). What is Co-Creation? Interaction Design Foundation—IxDF. The Interaction Design Foundation. https://www.interaction-design.org/literature/topics/co-creation

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Junklewitz, H., Hamon, R., André, A., Evas, T., Soler Garrido, J., & Sanchez Martin, J. I. (2023). Cybersecurity of artificial intelligence in the AI Act: Guiding principles to address the cybersecurity requirement for high risk AI systems. Publications Office of the European Union. https://data.europa.eu/doi/10.2760/271009

Leijnen, S., Aldewereld, H., van Belkom, R., Bijvank, R., & Ossewaarde, R. (2020). An agile framework for trustworthy AI. 75–78.

Miller, G. J. (2022). Stakeholder roles in artificial intelligence projects. Project Leadership and Society, 3, 100068. https://doi.org/10.1016/j.plas.2022.100068

Moore, S. (2023, December 18). What is...a Corporate Legal and Compliance team? | LinkedIn. https://www.linkedin.com/pulse/what-isa-corporate-legal-compliance-team-steven-moore-iczqf/

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. Science and Engineering Ethics, 26(4), 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

OECD. (2019). AI-Principles Overview. OECD AI Principles Overview. https://oecd.ai/en/principles

Peter, A. (2024, October 2). AI Software Architecture: Navigating the Essentials of Business App Development | LinkedIn. https://www.linkedin.com/pulse/ai-software-architecture-navigating-essentials-business-alexx-peter-uugnc/

Prem, E. (2023). From ethical AI frameworks to tools: A review of approaches. AI and Ethics, 3(3), 699–716. https://doi.org/10.1007/s43681-023-00258-9

Rochel, J., & Evéquoz, F. (2021). Getting into the engine room: A blueprint to investigate the shadowy steps of AI ethics. AI & SOCIETY, 36(2), 609–622. https://doi.org/10.1007/s00146-020-01069-w

Rodrigues, R. (2020). Legal and human rights issues of AI: Gaps, challenges and vulnerabilities. Journal of Responsible Technology, 4, 100005. https://doi.org/10.1016/j.jrt.2020.100005

Saltz, J. (2023, June 1). What is the AI Life Cycle? Data Science Process Alliance. https://www.datascience-pm.com/ai-lifecycle/

Sanin, A. (2020, September 4). Co-Creation How-To's. Design Globant. https://medium.com/design-globant/co-creation-how-tos-e25696f56d6f

UNESCO. (2023a). Ethical impact assessment. A tool of the Recommendation on the Ethics of Artificial Intelligence. UNESCO. https://doi.org/10.54678/YTSA7796

UNESCO. (2023b). Readiness assessment methodology. A tool of the Recommendation on the Ethics of Artificial Intelligence. UNESCO. https://doi.org/10.54678/YHAA4429

UNESCO. (2023c, April 21). Artificial Intelligence: Examples of ethical dilemmas | UNESCO. https://www.unesco.org/en/artificial-intelligence/recommendation-ethics/cases

White, C. (2023, November 30). What Does a UX Designer Actually Do? [2024 Guide]. https://careerfoundry.com/en/blog/ux-design/what-does-a-ux-designer-actually-do/

Zhou, J., & Chen, F. (2022). AI ethics: From principles to practice. AI & SOCIETY. https://doi.org/10.1007/s00146-022-01602-z

Zhou, J., Chen, F., Berry, A., Reed, M., Zhang, S., & Savage, S. (2020). A Survey on Ethical Principles of AI and Implementations. 2020 IEEE Symposium Series on Computational Intelligence (SSCI), 3010–3017. https://doi.org/10.1109/SSCI47803.2020.9308437

Zwikael, O., & Meredith, J. R. (2018). Who's who in the project zoo? The ten core project roles. International Journal of Operations & Production Management, 38(2), 474–492. https://doi.org/10.1108/IJOPM-05-2017-0274

CU2 | AI privacy and convenience

Index

1. Introduction

In the digital age, the rapid advancement of Artificial Intelligence (AI) has transformed countless aspects of daily life, reshaping how we interact with technology and each other. At the heart of this transformation lies the dynamic interplay between two critical factors: privacy and convenience. This CU delves into the complex relationship between these elements, exploring the challenges and opportunities they present in the era of smart technologies.

As we increasingly integrate AI into our personal and professional lives, it becomes imperative to critically examine the balance between enhancing user convenience and safeguarding personal privacy. This CU is designed to provide participants with a comprehensive understanding of how privacy and convenience coexist harmoniously within AI applications. It stresses highlighting both their potential to improve everyday life and the ethical considerations they entail.

Participants will gain insights into the fundamental principles governing data privacy, the ethical dilemmas posed by AI, and the legal frameworks shaping the future of technology deployment. By exploring real-world applications and hypothetical scenarios, this course aims to equip learners with the tools to navigate the evolving landscape of AI responsibly, ensuring that technological advancements enhance human experiences without compromising individual rights.

This journey through the realms of AI, privacy, and convenience will challenge participants to think critically about the role of technology in society and their responsibility as future leaders in AI.

IMAGE SOURCE | Generated by DALL-E

TIPS AND RECOMMENDATIONS FOR TEACHERS

An engaging question or a surprising fact to hook the learners' interest to explain the AI landscape today, emphasizing the significance of privacy and convenience.

Surprising Fact

A recent study found that AI systems can make decisions about your job applications, credit scores, and even your health care — often without your direct input or knowledge. This integration of AI raises crucial questions: How much control do we have over our data, and what does it mean for our privacy? Let’s dive into the complex world of AI, where convenience might come at the cost of our privacy.

Engaging Question

Did you know that the average person interacts with AI more than 20 times daily without realizing it? From personalized shopping recommendations to smart home devices managing our daily routines, AI is seamlessly integrated into our lives. But how often do you consider what personal information you're trading for this convenience? Let's explore what's really at stake.

Also, the content can be presented as a poll.

Poll Question

"Which services do you use that you know to collect your data to improve functionality and personalize your experience?"

Options

- Smartphones
- Social Media Platforms
- Streaming Services
- E-commerce Websites
- Smart Home Devices
- Fitness Trackers and Health Apps
- Voice Assistants
- Navigation and Travel Apps
- Banking and Financial Apps
- Email Services

Follow-Up Question

"How comfortable are you with these services collecting your data?"

Very comfortable || Somewhat comfortable || Neutral || Somewhat uncomfortable || Very uncomfortable

This approach helps identify awareness of data collection across different services and captures feelings about privacy and personal data use, providing a comprehensive view of participants' attitudes toward data privacy.

2. Sarah's dilemma: the privacy-convenience trade-off

IMAGE SOURCE | Generated by DALL-E

Imagine Sarah, a college student who just moved into her first apartment. Eager to streamline her busy life, Sarah invests in several smart devices: a smart speaker to manage her schedule and play music, a smart thermostat to control her apartment's temperature, and a smartwatch to monitor her fitness routines.

TIPS AND RECOMMENDATIONS FOR TEACHERS

Tip: ask questions at critical points to make it relatable

IMAGES SOURCE | Generated by DALL-E

One morning, Sarah's smart speaker suggests a new route to her college campus, saving her time during a heavy traffic period - a perfect example of convenience. Later, she gets a personalized ad on her phone for a coffee shop along her new route, tempting her with her favourite latte. Initially pleased, Sarah wonders,

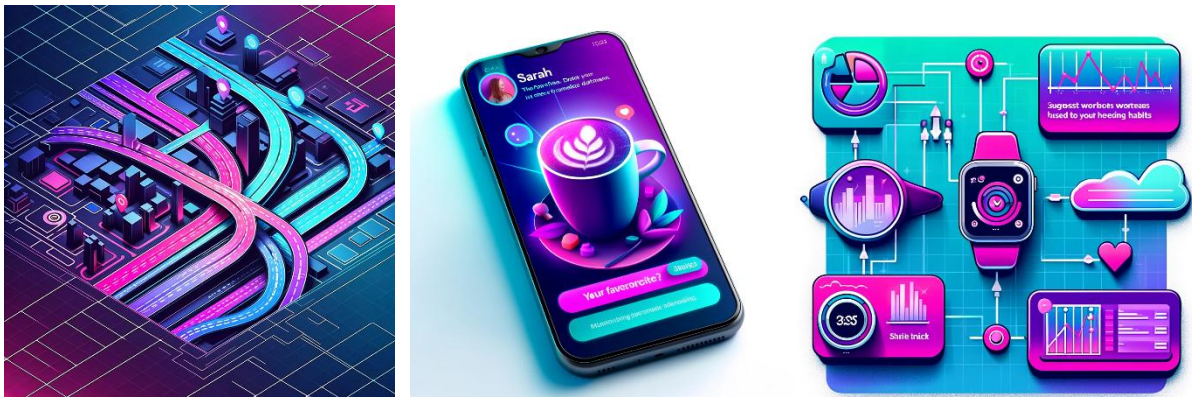
"How did it know?"

Question

"Do you think Sarah considered the privacy implications before purchasing these devices? Why or why not?"

Question

- How do you feel about devices making decisions based on your daily routines? What are the potential privacy trade-offs?
- What information might have been shared to personalize this advertisement? Are you comfortable with this level of data sharing for personalized ads?



IMAGES SOURCE | Generated by DALL-E

As days pass, Sarah observes more of these occurrences. Her smartwatch suggests exercise plans based on her fitness level and recent fast-food purchases, logged by her payment app. Her thermostat adjusts to her schedule and seems to anticipate unusual changes in her routine, like staying home when she's sick.

Question

"At what point does personalization become too invasive? Where would you draw the line for yourself?"

IMAGE SOURCE | Generated by DALL-E

While Sarah appreciates the convenience these AI-driven devices offer, making her life smoother and more efficient, she starts to feel uneasy about how much personal information they collect and how it's used. She didn't explicitly agree to share her location or her purchase history. Sarah's delight with the conveniences quickly becomes tinged with concern about her privacy.

This story introduces us to the delicate balance between privacy and convenience in AI. As AI technologies become more integrated into our everyday lives, they offer unprecedented convenience and raise significant privacy concerns. The challenge lies in finding a balance where Sarah and users like her can enjoy the benefits of AI without feeling that their personal information is being compromised.

IMAGE SOURCE | Generated by DALL-E

Question:

- Has technology ever made you feel similar to how Sarah feels? What actions did you take, if any?
- What would you advise Sarah to do in this situation?
- What steps can she take to control her data better?

3. Understanding Privacy in AI

Privacy can be broadly understood as the right of individuals to keep their personal information out of public view and control their data and interactions. This right is recognized as a fundamental human right in many legal systems and international treaties, emphasizing the importance of personal autonomy and dignity.

In the realm of AI, privacy takes on additional complexity. AI systems often rely on vast data to learn and make decisions. This data can include sensitive personal information, which, if mishandled, can lead to significant privacy breaches. Integrating AI in everyday technologies - from personal assistants and wearable devices to predictive analytics in healthcare - means that protecting privacy requires traditional data protection measures and new approaches to address the unique challenges posed by AI technologies.

"Privacy is a fundamental human right recognized in the UN Declaration of Human Rights, the International Covenant on Civil and Political Rights, and many other global and regional treaties. In the context of AI, privacy encompasses several key dimensions."

The significance of privacy in the context of AI is multifaceted:

- **Trust**: Effective privacy protections are crucial for maintaining user trust in AI technologies. Without trust, users may be reluctant to adopt potentially beneficial innovations.
- **Ethical Obligations**: There is an ethical imperative to respect and protect individual privacy. This involves ensuring that AI systems do not infringe on personal rights and are designed with privacy in mind from the ground up.
- **Compliance**: Many jurisdictions have stringent data protection regulations, such as the GDPR in Europe, which impose specific obligations on organizations that process personal data through AI systems.

3.1. Data Privacy: Protection of Personal Data from Unauthorized Access

Data privacy in AI is concerned with protecting personal data from unauthorized access, use, or disclosure. AI systems must be designed to ensure that data and susceptible information are handled securely and only accessed by authorized parties. Effective data privacy measures help prevent identity theft, financial fraud, and personal data exploitation risks. They include:

- **Data Encryption**: Encrypting data both at rest and in transit to prevent unauthorized access.
- **Access Controls**: Limiting data access to individuals who need it to perform their job functions.
- **Data Anonymization**: Removing personally identifiable information from datasets used in AI to reduce the risk of privacy breaches.

For instance, an online retail company collects customer data for order processing and personalized marketing. The company uses encryption to secure customer payment details and personal information to ensure data privacy. Additionally, they implement strict internal policies that limit employee access to this sensitive data only to those who need it to process orders. They also secure customer consent through clear opt-in procedures before sending marketing communications.

3.2. Information Security: Measures to Protect Data Integrity and Confidentiality

Information security is closely related to data privacy but focuses more on the protective measures that ensure data remains accurate, reliable, and accessible only to authorized users. In the context of AI:

Integrity: Protecting data from unauthorized changes that could compromise its accuracy.

Confidentiality: Ensuring that personal information is kept secret and only accessible to individuals who are authorized to access it.

Security Protocols: Implementing robust protocols such as secure socket layers (SSL), firewalls, and intrusion detection systems to protect against external attacks.

For instance, a healthcare provider uses an electronic health records (EHR) system to store patient data. To protect this data, the provider employs multiple layers of security measures. This includes using strong data storage and transfer encryption, multi-factor authentication (MFA) for system access, and regular security audits to identify and mitigate vulnerabilities. These practices help ensure the integrity and confidentiality of patient data against cyber threats.

3.3. Personal Autonomy: The Right to Control Information About Oneself

Personal autonomy in the digital age, especially within AI, is about maintaining control over personal information and the decisions made based on that information. AI systems must be designed to respect and enhance personal autonomy by:

- **Informed Consent**: Ensuring that users are fully informed about how their data will be used and obtaining their consent before collecting data.
- **Transparency**: Providing users with clear information about data processing practices and the logic behind AI decisions.
- **Control Mechanisms**: Provide users with mechanisms to control how their information is used and to opt out of data processing when they choose.

For instance, a fitness tracking app collects user physical activity data to provide personalized fitness advice. The app respects personal autonomy by clearly informing users about the types of data it collects and how it will be used. It also allows users to access their data, correct any inaccuracies, and opt out of data

collection for certain features. This approach ensures users have control over their personal information and its decisions.

4. Data Privacy

Data privacy refers to protecting personal data from unauthorized access, use, or disclosure. It encompasses ensuring that individuals have control over their information and that it is used in accordance with their preferences and legal regulations.

4.1. Importance of Data Privacy in AI

Data privacy is paramount in AI, where vast amounts of personal data are processed. Without adequate protection, individuals' sensitive information can be vulnerable to breaches and misuse, leading to potential harm, such as identity theft, financial loss, or discrimination. Moreover, AI systems rely heavily on data to make informed decisions, meaning that the quality and integrity of the data directly impact the performance and reliability of these systems.

4.2. Role of Regulations like GDPR

Regulations like the General Data Protection Regulation (GDPR) are crucial in enforcing data privacy principles in AI. GDPR sets strict requirements for collecting, processing, and storing personal data, including obtaining explicit consent from individuals, implementing appropriate security measures, and providing mechanisms for data subjects to access and control their data. By enforcing these regulations, GDPR aims to safeguard individuals' privacy rights and hold organizations accountable for handling personal data.

For instance, Health App Scenario: Consider a health app that collects and stores users' medical data, such as their medical history, diagnoses, and treatment plans. This app poses significant risks to users' data privacy without proper consent or security measures. For instance, if the app lacks robust encryption or access controls, it could be vulnerable to data breaches, exposing sensitive medical information to unauthorized parties. Moreover, if the app shares user data with third parties without explicit consent or anonymization, it could violate privacy regulations like GDPR, leading to legal consequences. This scenario illustrates the importance of implementing stringent data privacy measures in AI applications to protect individuals' sensitive information and ensure compliance with regulatory requirements.

4.3. Key Aspects of Data Privacy

Data privacy is critical to modern digital practices, especially with the proliferation of data-driven technologies and services. Here are some key elements of data privacy that help safeguard personal information and ensure the ethical handling of data:

- **Consent:** Consent involves obtaining permission from individuals before collecting, using, or sharing their data. It must be given freely, be specific, informed, and unambiguous. This means that individuals should be fully aware of what they are consenting to and clearly understand the scope and implications of the data processing activities.
- **Purpose Limitation:** This principle dictates that data should be collected for specified, explicit, and legitimate purposes and not further processed in an incompatible way. Purpose limitation helps ensure that data is used only in ways that the data subject expects and consents to, preventing data from being used maliciously or irresponsibly.
- **Data Minimization:** Data minimization means that only the data necessary for processing is collected and processed. This principle encourages organizations to limit data collection to what is directly relevant and necessary to accomplish a specified purpose. This reduces the risk of harm from data breaches, as less data can be compromised.
- **Accountability:** Accountability refers to organizations' obligation to take responsibility for the data they process and demonstrate compliance with all principles of data protection. This includes keeping records of data processing activities, implementing data protection measures, and showing how compliance is achieved. This principle ensures that organizations are transparent about their data handling practices and can be held responsible for their data processing activities.
- **Transparency:** Transparency requires that any information and communication relating to the processing of personal data be easily accessible and understandable to individuals. Privacy policies and data collection notices must use clear, plain language. Transparency builds trust, enhances fairness, and empowers users to make informed decisions about their data.
- **Access and Correction:** Individuals have the right to access their data and to have incorrect data about them corrected. Providing subjects with the ability to access and update their data as needed ensures data accuracy and allows individuals to maintain control over their information, which is fundamental to personal autonomy.
- **Security:** Security involves implementing appropriate technical and organizational measures to protect personal data against accidental or unlawful destruction, loss, alteration, unauthorized disclosure, or access to

it. This includes practices like encryption, secure data storage, and controlled physical and digital access.

Together, these aspects of data privacy form a robust framework that organizations can use to protect personal data and respect the privacy rights of individuals. They help build a privacy-conscious culture that aligns with legal standards and ethical expectations in the digital age.

5. Understanding Convenience in AI

In the context of AI, convenience refers to the ease and efficiency with which tasks can be completed or experiences enhanced through AI technologies. AI streamlines processes reduce manual effort and provides personalized experiences tailored to individual preferences.

Convenience, facilitated by AI technologies, encompasses the seamless integration of automation, personalization, and efficiency across various domains. From healthcare and finance to everyday life, AI-driven solutions optimize processes, empower decision-making, and enrich experiences, ultimately enhancing productivity and improving overall well-being.

Convenience, in the realm of AI, embodies the seamless integration of technology to streamline tasks and amplify efficiency. AI technologies leverage algorithms and machine learning to automate processes, reduce manual effort, and tailor experiences to individual preferences. AI systems optimize workflows by analysing vast amounts of data and learning from patterns, saving time and resources while delivering enhanced outcomes. Below are a few points where AI contributes to convenience in several areas:

- **Automation of Routine Tasks:** AI excels at automating repetitive and mundane tasks that traditionally require human intervention. For instance, AI-driven software can handle scheduling appointments, responding to customer inquiries, or managing emails automatically. This speeds up the process and frees up human resources for more complex and creative tasks, thereby increasing productivity and efficiency.
- **Personalization:** AI systems can analyse vast amounts of data to tailor services and products to individual preferences and needs. In consumer contexts, this might manifest as AI recommending products based on past purchases or viewing habits, as seen in online retail and streaming services. In more personal applications, AI can adjust home heating systems based on the homeowner's schedule and preferences, ensuring comfort while optimizing energy use.
- **Enhanced Decision Making:** AI enhances decision-making by providing individuals and organizations with insights derived from analysing large datasets that would be too complex for humans to process manually. In sectors like healthcare, AI algorithms assist in diagnosing diseases by

analysing medical imaging data faster and often more accurately than human practitioners. In finance, AI systems can analyse market data to provide investment insights or detect fraudulent transactions with high accuracy and speed.

- **Efficiency and Resource Management:** AI significantly improves overall efficiency and optimizes resource management. For instance, AI-driven logistics and supply chain management systems can rapidly predict demand, optimize delivery routes, and manage inventory, reducing waste and costs. Similarly, smart grids powered by AI can manage electricity distribution across a city more efficiently, adapting to changes in demand in real time.
- **Accessibility and Ease of Use:** AI technologies often make technology more accessible and easier for a broader range of people, including those with disabilities. Voice-assisted devices powered by AI help individuals with vision impairments or physical limitations interact with technology and manage their environment more effectively. AI can also break down language barriers with real-time translation services, making information and communication accessible to non-native speakers.

5.1. Real-life Examples of Convenience

Healthcare: In healthcare, AI technologies offer convenience by automating diagnostic processes, analysing medical images, and detecting patterns in patient data to assist healthcare professionals in diagnosing diseases more quickly and accurately. This saves healthcare providers and patients time, leading to more efficient healthcare delivery and improved patient outcomes.

IMAGE SOURCE | Generated by DALL-E

Convenience in Healthcare through AI

- **Automated Diagnostic Processes**: AI speeds up medical data analysis, such as imaging and diagnostic tests, allowing for quicker diagnoses.
- **Enhanced Diagnostic Accuracy**: AI algorithms detect subtle patterns in data that humans might miss, increasing the accuracy and reliability of diagnoses.
- **Predictive Analytics**: AI uses data from various sources, including wearable technology, to predict potential health issues before they become severe, facilitating preventive healthcare.
- **Streamlined Patient Management**: AI optimizes hospital workflows by managing schedules, patient intake, and routing to appropriate care settings, reducing wait times and improving patient experiences.

- **Personalized Treatment Plans**: Algorithms analyse individual patient data to tailor treatment plans that are most likely to be effective based on the patient's unique health profile.

Finance: AI-powered financial tools provide convenience by offering personalized investment advice based on individual financial goals, risk tolerance, and market trends. These tools use algorithms to analyse vast amounts of financial data and provide recommendations for investment strategies, helping individuals make informed decisions and optimize their financial portfolios.

IMAGE SOURCE | Generated by DALL-E

Convenience in Finance through AI

- **Personalized Investment Advice**: AI tools analyse an individual's financial goals, risk tolerance, and personal economic situations to provide tailored investment recommendations.
- **Advanced Data Analysis**: These systems utilize algorithms to sift through large volumes of financial data, including market trends and historical performance, to identify the best investment opportunities.
- **Optimized Portfolio Management**: AI helps dynamically adjust investment portfolios based on real-time market changes and individual financial goals, enhancing potential returns while managing risk exposure.
- **Efficient Market Insights**: AI tools provide quick insights into market conditions, helping investors understand complex market dynamics and make timely decisions.
- **Risk Assessment**: Algorithms evaluate the risk associated with various investment options, allowing users to make informed choices based on risk tolerance.
- **Automation of Transactions**: AI can automate trading and investment transactions, streamlining the execution process and reducing the likelihood of human error.
- **Enhanced Security**: Advanced security measures, like anomaly detection, are employed by AI to identify and mitigate potential threats to investment accounts.

Everyday life: AI-enabled smart home devices, such as voice assistants, smart thermostats, and automated home security systems, offer convenience by automating household tasks and enhancing comfort and security. These devices learn user preferences over time and adjust settings, accordingly, providing seamless and personalized experiences that simplify daily routines and improve quality of life.

IMAGE SOURCE | Generated by DALL-E

Convenience in Everyday Life through AI

- **Automation of Household Tasks**: AI-enabled devices like smart vacuum cleaners and automated kitchen appliances handle routine chores, freeing up users' time.
- **Personalized Settings**: Devices such as smart thermostats and lighting systems learn and adapt to user preferences, automatically adjusting settings for optimal comfort.
- **Voice-Assisted Technology**: Voice assistants like Alexa and Google Assistant enable hands-free control over household devices, making information access and task management easier.
- **Enhanced Home Security**: Smart security systems use AI to monitor home environments, recognize familiar faces, detect unusual activities, and alert homeowners to potential threats.
- **Energy Efficiency**: AI-driven devices optimize energy use in the home by learning peak usage times and adjusting operations to reduce waste, lowering utility bills.
- **Seamless Integration**: Smart home ecosystems connect various devices, allowing them to work together smoothly, enhancing user convenience through integrated controls and interactions.
- **Remote Monitoring and Control**: Homeowners can monitor and control smart home devices remotely via smartphones, providing peace of mind when away from home.

While AI significantly enhances convenience, it also raises important considerations, such as privacy concerns, the need for human oversight, and the potential for job displacement in specific sectors. Ensuring that AI systems are used ethically and responsibly is crucial to maximizing their benefits while minimizing potential harm. In conclusion, convenience in AI is about harnessing the power of artificial intelligence to make life easier, more enjoyable, and more efficient. As AI technologies evolve and integrate into various aspects of life, their

potential to drive convenience will likely expand, promising transformative impacts across all facets of society.

TIPS AND RECOMMENDATIONS FOR TEACHERS

Present hypothetical scenarios that challenge participants to decide between prioritizing privacy or convenience.

Example

Hypothetical Scenario: Health Tracker App

Background: You're considering a new health tracker app, HealthMate. This app offers personalized fitness and diet plans by analysing your daily activities, food intake, and health metrics.

Scenario Challenge: HealthMate can sync with your social media to gather more lifestyle data for enhanced personalization and share insights with partners who might send you targeted ads based on your health goals.

Questions for Reflection:

- Would you opt into the enhanced personalization, knowing your data might be used for targeted advertising?
- How do you weigh the benefits of a customized health plan against the risks of sharing your data?

This scenario prompts participants to evaluate the balance between the convenience of tailored health recommendations and the privacy risks of data sharing.

6. The Interplay of Privacy and Convenience

The dynamic interplay between privacy and convenience in modern technology presents opportunities and challenges. As technology increasingly integrates into our daily lives, understanding how to balance these aspects becomes crucial. Below are the implications of the trade-offs, ethical dilemmas, and real-world consequences of balancing privacy with convenience.

Trade-offs Between Privacy and Convenience:

Balancing privacy and convenience requires careful navigation of the inherent trade-offs involved:

- Restrictions on Data Collection: Implementing privacy protections often means limiting the amount and type of data collected and used. This can restrict the functionality of AI systems that rely on large datasets to improve service delivery and user experience.
- Sacrificing Privacy for Convenience: On the flip side, enhancing convenience often involves collecting and analysing extensive personal data. This can include tracking user behaviours, preferences, and locations, which might encroach on personal privacy.

Ethical Dilemmas Arising from These Trade-offs:

The balance between privacy and convenience brings several ethical dilemmas to the forefront:

- Compromise of Privacy: How much privacy are individuals willing to forfeit for enhanced convenience? The answer often varies based on individual perceptions and the context of the technology's use.
- Data Collection Practices: Organizations are also ethically responsible for collecting, using, and sharing personal information. Ethical considerations must guide decisions about what data is collected, how it is used, and how much users are informed about these practices.

Real-world Implications of Choosing Convenience Over Privacy or Vice Versa:

The decisions made regarding the balance between privacy and convenience have tangible effects:

- Consequences of Prioritizing Convenience: Choosing convenience often increases risks such as data breaches, identity theft, and significant privacy invasions. This can lead to long-term damage to user trust and potential legal repercussions for companies.
- Effects of Prioritizing Privacy: Conversely, heavily prioritizing privacy might limit the effectiveness of specific technological solutions. This could lead to less personalized services and potentially slow down technological innovation and advancement that could benefit society.

Navigating the balance between privacy and convenience requires a nuanced approach that considers technology's ethical, social, and personal implications. Organizations and individuals must strive to find a middle ground that respects user privacy while also taking advantage of the benefits that technology can bring. This balance is not static and needs continuous evaluation as technology evolves and societal values shift. Understanding these dynamics helps stakeholders make informed decisions that protect individual rights while embracing the advancements of the digital age.

7. Importance of Privacy and Convenience

Privacy and convenience in AI are pivotal in shaping user trust and adoption of technology. As AI integrates into our daily lives, achieving a balance between protecting personal data and enhancing user experiences is crucial. Privacy ensures safeguarding personal information, while convenience enhances efficiency and personalization in technology use. The challenge lies in harmonizing these aspects to foster ethical AI deployment and societal acceptance, ensuring that advancements in AI deliver benefits without compromising individual rights.

Enhancing User Trust and Technology Adoption

Prioritizing privacy and convenience in AI technologies is essential for building user trust and encouraging broader adoption. When AI systems transparently manage privacy and provide seamlessly convenient experiences, they are more readily accepted by users. Here's how prioritizing these aspects can influence user trust and technology adoption:

- **Building Trust**: Users tend to trust technology that respects their privacy and enhances their daily lives without significant intrusion. AI solutions that effectively protect user data while offering personalized and efficient experiences tend to build a positive reputation among users.
- **Encouraging Adoption**: When users perceive that an AI technology enhances their productivity or convenience without compromising their privacy, they are more likely to adopt and recommend it. Demonstrating the value of AI through convenience, coupled with robust privacy safeguards, motivates users to integrate these technologies into their daily routines.

Legal and Ethical Considerations in AI Deployment

Deploying AI technologies requires careful consideration of both legal and ethical implications to ensure responsible use and compliance with regulations:

- **Regulatory Compliance**: Laws such as the General Data Protection Regulation (GDPR) provide strict guidelines on data privacy that AI technologies operating in or targeting users in the European Union must follow. Compliance with these regulations is not just about legal necessity but also about demonstrating a commitment to user privacy.
- **Ethical Frameworks**: Beyond legal requirements, ethical frameworks guide AI development to ensure that technologies are used for the benefit of society. These frameworks help developers and companies navigate complex issues such as bias, fairness, transparency, and accountability in AI systems.

Impact on Societal Norms and Individual Behaviours

AI's role in balancing privacy and convenience has significant effects on societal norms and individual behaviours:

- **Shaping Privacy Expectations and Preferences**: As AI technologies become more embedded in everyday life, they influence how individuals perceive and value their privacy. This can lead to changing expectations, with some users becoming more privacy-conscious while others may become desensitized to sharing personal information.
- **Influencing Daily Routines and Decision-Making**: AI technologies that offer convenience can significantly alter daily life. Smart home devices, personalized shopping experiences, and automated personal finance management can reshape everyday activities, making life easier and creating dependencies on technology for decision-making.

By understanding and addressing the complex interplay between privacy and convenience, AI developers and deployers can enhance the technology's acceptance and ensure it contributes positively to societal development and individual well-being. These considerations are crucial for the sustainable growth of AI technologies in a society increasingly conscious of privacy and ethical standards.

8. Navigating Risks and Benefits

Potential risks associated with AI technologies under privacy and convenience include:

- **Data breaches:** AI systems often handle vast amounts of sensitive data, making them targets for malicious attacks. Data breaches can result in unauthorized access to personal information, leading to identity theft, financial loss, and reputational damage.
- **Loss of privacy:** AI technologies may collect and analyse personal data without individuals' consent or knowledge, compromising privacy rights. This can erode trust between users and AI systems, leading to data misuse and exploitation concerns.
- **Biases in AI algorithms:** AI algorithms may perpetuate biases in training data, leading to discriminatory outcomes. Biased AI algorithms can result in unfair treatment or decision-making, exacerbating societal inequalities and undermining trust in AI systems.

Consequences of these Risks:

These risks can undermine trust in AI systems and lead to negative consequences for individuals and society:

- **Erosion of trust:** Data breaches and privacy violations can erode trust between users and AI systems, leading to decreased adoption and usage. Individuals may become reluctant to share their data or engage with AI technologies, hindering their potential benefits.
- **Negative societal impacts:** Biased AI algorithms can perpetuate discrimination and inequality, leading to unfair treatment and social unrest. This can damage social cohesion and trust in institutions, exacerbating societal divides.
- **Legal and regulatory repercussions:** Data breaches and privacy violations can result in legal and regulatory repercussions for organizations responsible for AI systems. Fines, lawsuits, and reputational damage can have significant financial and operational implications.

Potential Benefits:

Despite the risks, AI technologies offer numerous potential benefits:

- **Increased Efficiency:** AI streamlines processes and automates tasks, increasing efficiency and productivity. By handling repetitive tasks, AI frees up time for employees to focus on higher-value activities.
- **Better Data-Driven Decisions:** AI algorithms analyse vast amounts of data to uncover insights and patterns, enabling organizations to make better-informed decisions. AI-driven insights can optimize operations, improve customer experiences, and drive innovation.
- **Enhanced User Experiences:** AI personalization enhances user experiences by delivering tailored recommendations and services. From personalized product recommendations to predictive maintenance in manufacturing, AI enhances user satisfaction and engagement.

Mitigating Risks:

Organizations should implement robust security measures to mitigate these risks, ensure transparency and accountability in AI systems, and prioritize ethical considerations throughout the AI development lifecycle. By proactively addressing these risks, organizations can build trust in AI systems and minimize potential harm to individuals and society.

To navigate the risks and benefits of AI technologies effectively, organizations can adopt the following strategies:

- **Implement Robust Security Measures:** Organizations should implement robust security measures to protect against data breaches and unauthorized access. This includes encryption, access controls, and regular security audits.
- **Ensure Transparency and Accountability:** Transparency and accountability are essential for building trust in AI systems. Organizations

should be transparent about how AI technologies are used and ensure accountability for their actions.

- **Prioritize Ethical Considerations:** Ethical considerations should guide AI development and deployment. Organizations should prioritize fairness, transparency, and accountability in AI systems to mitigate biases and ensure ethical use of data.

By adopting proactive risk management and responsible AI deployment practices, organizations can maximize the benefits of AI technologies while minimizing potential harm to individuals and society.

9. Case study: AI in surveillance

Scenario

Local authorities implement AI-powered surveillance systems in a bustling city to enhance public safety and security. These systems utilize advanced technologies such as facial recognition, behaviour analysis, and object detection to monitor public spaces, identify potential threats, and respond to emergencies in real time.

IMAGE SOURCE | Generated by DALL-E

Security Enhancements vs. Privacy Breaches:

Security Enhancements	Privacy Breaches
The AI surveillance systems bolster security measures by enabling law enforcement agencies to quickly detect and respond to suspicious activities or potential security threats. Facial recognition technology helps identify individuals on watchlists or those reported missing, facilitating timely interventions and enhancing public safety.	Despite the security benefits, concerns arise regarding the potential privacy breaches associated with AI surveillance. Continuous monitoring and data collection in public spaces raise questions about individuals' right to privacy and freedom from constant surveillance. There is a risk of misidentification and false positives, leading to wrongful accusations and invasions of privacy for innocent individuals mistakenly flagged by the system.

Importance of Thoughtful AI Implementation:

The case of AI in surveillance highlights the importance of thoughtful AI implementation to balance security needs with privacy considerations:

- **Clear Policies and Regulations:** Thoughtful AI implementation involves establishing clear policies and regulations governing surveillance technologies to ensure transparency, accountability, and adherence to privacy laws.
- **Ethical Use of Data:** Organizations must prioritize the ethical use of data collected through surveillance systems, respect individuals' rights to privacy, and minimize the risk of privacy breaches or misuse of personal information.
- **Algorithmic Transparency and Accountability:** Implementing transparent AI algorithms and mechanisms for accountability ensures that surveillance systems are fair, accurate, and accountable for their actions, mitigating the risk of biases or errors that could lead to privacy violations.
- **Public Engagement and Oversight:** Engaging the public in discussions about AI surveillance and involving stakeholders in decision-making

promotes transparency, trust, and accountability, fostering responsible deployment and usage of surveillance technologies.

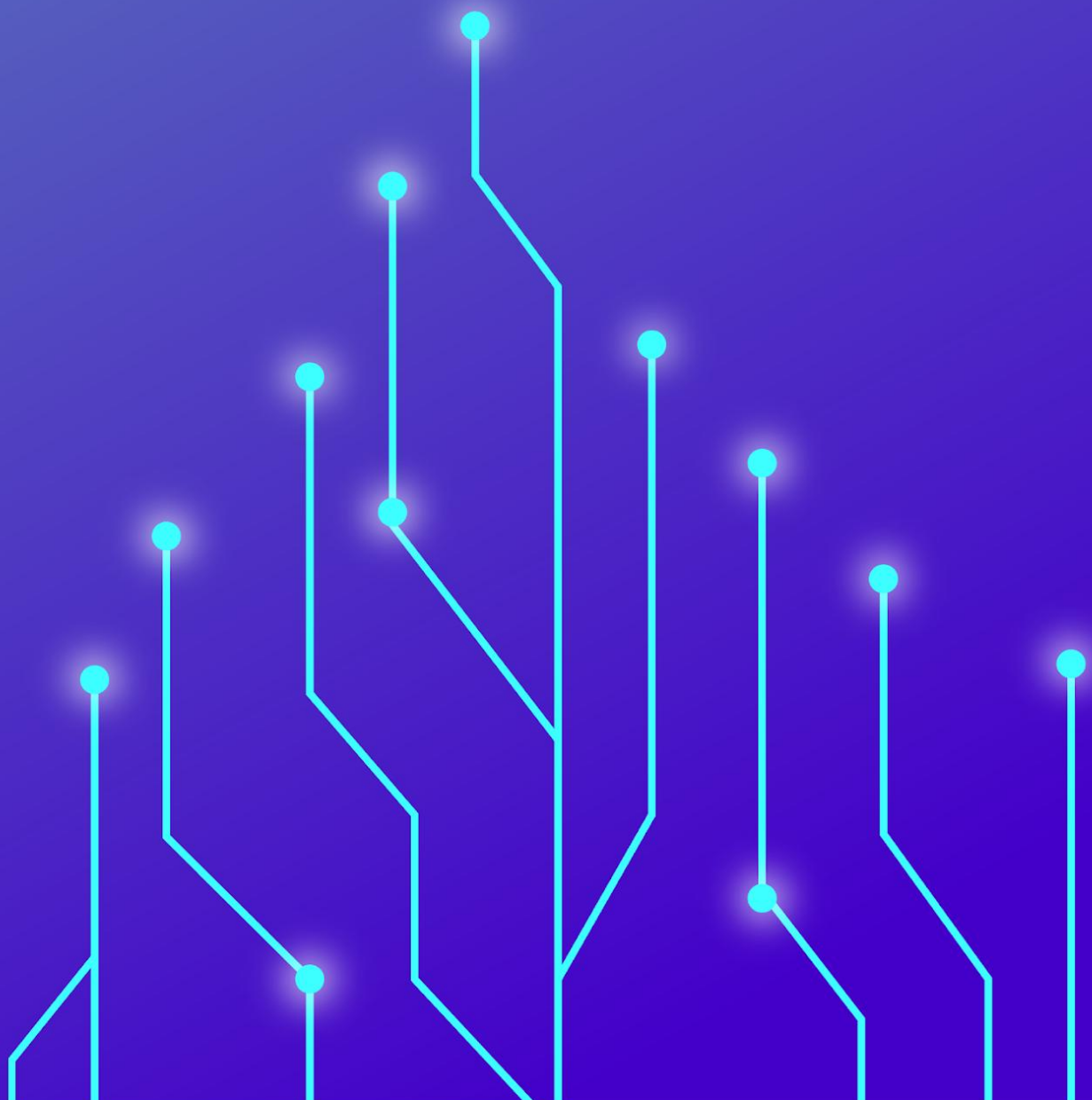
In conclusion, the case of AI in surveillance underscores the need for thoughtful AI implementation that balances security enhancements with privacy considerations. Organizations can deploy surveillance systems that enhance security while prioritizing ethical principles, transparency, and accountability, respecting individuals' privacy rights, and fostering public trust.

10. Conclusion

Exploring privacy and convenience within artificial intelligence (AI) highlights a critical balance necessary for fostering trust and broader technology adoption. As AI permeates various aspects of healthcare, finance, and daily life, it brings tremendous benefits like improved efficiency and personalized services, accompanied by the need for robust privacy protections and ethical considerations. Adhering to legal frameworks such as GDPR and addressing ethical dilemmas is essential to ensure responsible AI deployment.

As AI shapes societal norms and individual behaviours, maintaining a balance between user convenience and data privacy becomes crucial. Developers, users, and policymakers must collaborate to ensure AI technologies enhance lives without compromising fundamental rights. Emphasizing transparency and accountability will be key to harnessing AI's potential responsibly, ensuring it remains a beneficial and trusted technology in our increasingly digital world.

CU3 | Algorithms and their limitations



Index

1. Introduction[1]

This handbook is designed to be a key tool for trainers delivering the Charlie Project training course. It will not only describe the basic contents of the course but also provide detailed instructions, tips, and recommendations to the trainers on how to effectively deliver the course. Additionally, the handbook will propose a variety of activities aimed at engaging trainees and enhancing their learning experience. It also includes guidelines on how to evaluate the learning outcomes of the trainees, ensuring that the course objectives are met and that both trainers and trainees have a clear understanding of the expected competencies to be developed throughout the training. This comprehensive approach will aid trainers in facilitating a dynamic and effective learning environment[2].

This unit will introduce the fundamental concepts of algorithms, their models, and limitations. Participants will gain an understanding of the role algorithms play in various sectors and the potential pitfalls arising from their use. Topics covered will include algorithmic complexity, optimization, and the limitations of algorithmic decision-making.

In today's digital age, algorithms play an increasingly significant role in shaping various aspects of our lives, from the content we consume online to the decisions made by automated systems. Understanding the intricacies of algorithmic complexity, optimization strategies, and the limitations inherent in algorithmic decision-making processes is crucial for anyone navigating this landscape.

Understanding **algorithmic complexity** entails examining the efficiency and performance of algorithms as they address diverse computational tasks. It involves comprehending the resources—such as time and memory—that algorithms necessitate to solve problems and how these requirements scale with larger input sizes. This comprehension empowers us to assess the viability and applicability of employing specific algorithms in real-world contexts (Cormen et al., 2009).

Optimization methodologies constitutes another fundamental aspect of comprehending algorithms. These methods aim to refine the efficiency and efficacy of algorithms by adjusting their parameters or restructuring their

[1] Note: In the preparation of this document, the authors employed Large Language Models (LLM) to assist with text proofreading, content structuring, and the identification of pertinent examples included in the material. After utilizing this AI tool, the authors meticulously reviewed and refined the content as needed. The authors take full responsibility for the content's integrity and accuracy in the final publication.

[2] To teacher/Trainer: This handbook is aimed to guide you through the material for the course unit. The support PPTs contain similar information to what the E-learning covered, but in more detailed form. The PPTs are extensive and are not meant to be covered in their entirety during the interactive sessions. Instead, you can ask at the beginning of the interactive sessions which topics were most difficult to comprehend for the students and use only those support slides during the session. You can ask this by using tools like Mentimeter or Kahoot as an example, adding the course topics from the competence unit content described in bullet points at the end of the introduction section. Furthermore, CU-based case study is aimed to be handled also during the interactive session. Any further suggestions in the handbook are only guidance; feel free to use it as you see fit.

fundamental logic. Whether it involves reducing computation time, maximizing resource utilization, or optimizing for particular objectives, employing appropriate optimization strategies can significantly enhance algorithm performance across various domains (Cormen et al., 2009).

However, alongside the potential benefits of algorithmic decision-making, there exist inherent limitations that demand careful consideration. Algorithms operate within predefined frameworks and are reliant on the quality of input data and the accuracy of underlying assumptions. Biases in data, flawed modelling assumptions, or unexpected edge cases can lead to erroneous outcomes or unintended consequences, highlighting the fragility of algorithmic decision-making systems.

Deepening one's comprehension of algorithmic complexity, optimization strategies, and the limitations inherent in algorithmic decision-making processes can foster a more robust understanding of algorithmic nuances and challenges. It equips individuals with the critical thinking skills necessary to evaluate algorithmic solutions critically, identify potential pitfalls, and advocate for responsible and ethical algorithmic practices in an increasingly algorithm-driven world.

TIPS AND RECOMMENDATIONS FOR TEACHERS

Tips and recommendations for teachers to engage with initial algorithm concepts

1. Clarify Key Concepts
Before diving deep into discussions, ensure that all participants have a clear understanding of basic terminology such as "algorithmic complexity," "optimization," and "limitations of algorithms." This can be achieved through a brief presentation or a glossary handout at the beginning of the session.

2. Relate to Real-world Applications
Connect abstract concepts to real-world applications in various sectors such as finance, healthcare, or e-commerce. This helps participants see the relevance and impacts of algorithms in everyday scenarios, making the content more engaging and understandable.

3. Use Visual Aids
Algorithms can be complex and abstract. Utilize diagrams, flowcharts, and animations to visualize concepts like algorithmic flow, complexity, and optimization processes. This can help participants more easily grasp and remember the information.

4. Encourage Questions
Create an open environment where participants feel comfortable asking questions. This can be facilitated through regular Q&A sessions after covering

each major topic, ensuring that participants fully comprehend the material before moving on.

5. Incorporate Case Studies
Discuss real cases where algorithmic decision-making has succeeded or faced challenges. Analysing these cases can spark discussion about the ethical implications, efficiency, and practical considerations of using algorithms.

6. Promote Critical Thinking
Encourage participants to think critically about the limitations and potential biases in algorithmic models. This could be through group discussions or debates on how to mitigate these issues in algorithm design and implementation.

Recommended warm-up activities

1. Algorithm Role-Play
A role-play activity where participants simulate an algorithm in action, either by manually sorting items, following a simple set of instructions to achieve a goal, or by acting out the steps an algorithm might take in decision-making processes. This exercise helps illustrate the concept of step-by-step processing in algorithms.

2. Algorithmic Complexity Icebreaker
Start with a simple task, like organizing books by colour, and scale it up to organizing by genre, then by author within each genre. Discuss how the complexity increases with each added layer, drawing parallels to algorithmic complexity and how resources are used more as tasks become more complex.

3. Optimization Challenge
Give small groups a set task, such as arranging chairs in a room to fit as many people as possible but with the fewest rows. Let them attempt it, discuss their strategies, and then reveal how optimization strategies could enhance their solutions.

4. Ethical Debate
Organize a debate on the ethical implications of algorithms in decision-making, focusing on potential biases and the consequences of relying heavily on automated systems. This can warm participants up to the complexities and responsibilities of working with algorithms.

5. Break-the-Algorithm Game
Provide participants with a simple algorithm or set of rules and challenge them to find the edge cases or scenarios where the algorithm fails. This activity can highlight the limitations and potential flaws in algorithmic logic.
These activities and strategies will not only make the learning process engaging but also empower participants with a deeper understanding and appreciation of the complexities involved in algorithmic decision-making.

2. Understand the fundamental concepts of algorithms, their models, and limitations

2.1. Algorithmic Complexity

Algorithms are nowadays ubiquitous. They guide computers to process information, solve problems and make important decisions. They can solve problems from organizing data to finding the best travel routes or suggesting movies. In an age dominated by big data and social media, knowing how algorithms work is key to understand their limitations and possible biases. They don't just improve how computers perform tasks; they also influence our everyday experiences, making them essential for modern computing.

An algorithm is a sequence of unambiguous instructions for solving a problem, it must be correct, always gives a correct solution, and it must be finite, must terminate. It's a step-by-step procedure that tells us what to do to achieve a certain goal. We can think of it like cooking: we follow a recipe that tells us exactly what ingredients to use, how much of each, and what steps to take. Similarly, an algorithm guides a computer or a person through a series of actions to accomplish a specific outcome, whether it's sorting a list of numbers, finding the shortest route on a map, or recommending movies based on your preferences.

The concept of algorithms has been around for a long time, dating back to ancient civilizations. One of the earliest examples is the Euclidean algorithm, attributed to the ancient Greek mathematician Euclid, which was developed around 300 BCE for finding the greatest common divisor of two numbers.

Throughout history, various cultures and civilizations developed their own algorithms to solve mathematical problems, navigate geographical terrain, or perform tasks efficiently. These algorithms were often passed down orally or through written texts.

The term "algorithm" itself comes from the name of the Persian mathematician Muhammad ibn Musa al-Khwarizmi, who lived during the Islamic Golden Age in the 9th century. Al-Khwarizmi wrote a book called "Al-Kitab al-Mukhtasar fi Hisab al-Jabr wal-Muqabala" (The Compendious Book on Calculation by Completion and Balancing), which introduced systematic methods for solving mathematical equations. The word "algorithm" is derived from the Latinized version of his name, "Algoritmi".

Since then, algorithms have evolved and become fundamental in various fields, especially with the advent of modern computing. Today, algorithms are used not only in mathematics but also in computer science, engineering, biology, economics, and many other disciplines to solve complex problems and automate tasks.

2.2. Algorithm characterization

Algorithm characterizations are attempts to formalize the word algorithm. The term algorithm does not have a generally accepted formal definition. Knuth (1997) defined a list of five properties that are widely accepted as requirements for an algorithm:

- **Finiteness:** "An algorithm must always terminate after a finite number of steps ... a very finite number, a reasonable number". Finiteness ensures that algorithms terminate after a finite number of steps, avoiding infinite loops that could consume computational resources indefinitely. It's akin to completing a puzzle where eventually, the final piece falls into place, signalling the algorithm's conclusion. This characteristic of finiteness provides predictability and control over computational processes.
- **Definiteness:** "Each step of an algorithm must be precisely defined; the actions to be carried out must be rigorously and unambiguously specified for each case". It emphasizes the need for unambiguous steps that leave no room for interpretation. A definite algorithm clearly outlines each action to be taken at every stage of problem-solving, ensuring that there is no ambiguity or confusion about what needs to be done. This characteristic is crucial because it enables anyone, regardless of their background or expertise, to understand and implement the algorithm correctly.
- **Input:** "...quantities which are given to it initially before the algorithm begins. These inputs are taken from specified sets of objects". Input refers to the data or information that the algorithm operates on to produce a result. This input could be any form of data, such as numbers, text, images, or even other algorithms. Input is essential because it provides the raw material that the algorithm processes and manipulates to solve a problem or perform a task.
- **Output:** "...quantities which have a specified relation to the inputs". Output refers to the result or solution produced by the algorithm after processing the input. This output could also take various forms depending on the nature of the problem being solved, such as a single value, a set of values, a decision, a visual representation, or any other form of information that represents the desired outcome.
- **Effectiveness:** "... all of the operations to be performed in the algorithm must be sufficiently basic that they can in principle be done exactly and in a finite length of time by a man using paper and pencil". Effectiveness emphasizes the practicality and feasibility of algorithms. Just as a recipe with obscure or unobtainable ingredients becomes ineffective, algorithms must utilize accessible resources to accomplish their tasks. By ensuring that algorithms are practical and feasible, effectiveness ensures that solutions are achievable within the constraints of available resources and time.

As we can observe, the above description of an algorithm characteristics may be intuitively clear, it lacks formal rigor, since it is not exactly clear what "precisely defined" means, or "rigorously and unambiguously specified" means, or "sufficiently basic", and so forth. We well consider it sufficient for our purposes.

2.3. Algorithm structure

One of the main concepts behind algorithms is the abstraction. Abstraction in this context refers to the idea of hiding complex details and focusing only on the essential aspects necessary for understanding and solving a problem. It's like zooming out to see the bigger picture without getting lost in the fine-grained details. In algorithm design, abstraction helps in simplifying the problem-solving process by breaking it down into smaller, manageable parts. This allows us to tackle each part separately, without needing to understand every intricate detail all at once.

Algorithms typically involve multiple instructions working together, rather than just a single step. When creating complex programs, we execute a series of these instructions in sequence, repeat them, or make choices based on specific conditions.

Knowing how algorithms work involves understanding different ways instructions can be combined. These schemes provide a structured approach to organizing and sequencing the instructions in an algorithm. They help in breaking down complex problems into manageable tasks, making the algorithm more understandable and easier to implement. Algorithms are typically composed of simple structures that are organized in a hierarchical manner. These structures control the flow of instructions within the algorithm, dictating the sequence in which them are executed. This is often referred to as the control structure of the algorithm.

A control structure essentially directs the order of execution of the instructions in an algorithm. It determines the path that the execution process takes, based on the conditions defined within the structure. Control structures are a fundamental aspect of any algorithm, as they ensure that the instructions are executed in the correct order and that the desired outcome is achieved.

One common feature of all control structures is that they have a single-entry point and a single exit point. The single-entry point marks the beginning of the control structure, where the execution process enters the structure. On the other hand, the single exit point signifies the end of the control structure, where the execution process leaves the structure and moves on to the next part of the algorithm. This feature ensures that the execution process follows a defined path within the control structure, maintaining the integrity and correctness of the algorithm.

There are three instruction composition schemes:

· **Sequential:** In this scheme, instructions are executed one after the other, following a linear order. It's very close to following a recipe step-by-step. Sequential algorithms are prevalent in everyday tasks, such as making a sandwich or calculating the sum of numbers. They ensure that actions occur in a specific sequence, without any branching or decision-making.

Example: An algorithm to make a cake.
1. Preheat the oven.
2. Grease and flour a cake pan.
3. In a mixing bowl, cream together butter and sugar.
4. Beat in eggs, one at a time.
5. Stir in vanilla extract.
6. In a separate bowl, combine flour, baking powder, and salt.
7. Gradually add the dry ingredients to the wet mixture, alternating with milk.
8. Pour the batter into the prepared cake pan.
9. Place the cake pan in the preheated oven.
10. Bake for 25-30 minutes, or until a toothpick inserted into the center comes out clean
11. Remove the cake from the oven and let it cool in the pan for 10 minutes.
15. Serve the cake.

· **Conditional or Alternative:** These instructions introduce decision points. Based on certain conditions, the program takes different paths. Think of it as choosing between multiple options. Conditional statements (like "if," "else," and "switch") allow us to handle various scenarios.

Example: Algorithm that simulates the process of deciding whether to bring an umbrella or a jacket when leaving the house based on the weather forecast:

1. Check the weather forecast for rain.
2. If the forecast predicts rain, then:
3. Take an umbrella.
4. Else If the forecast does not predict rain and temperature is below 15 degrees Celsius then:
5. Take a jacket.
6. Leave the house.

Note that the indentation of instructions 3 and 5 has a specific role in the algorithm definition. These instructions only are executed if the previous instruction is evaluated positively.

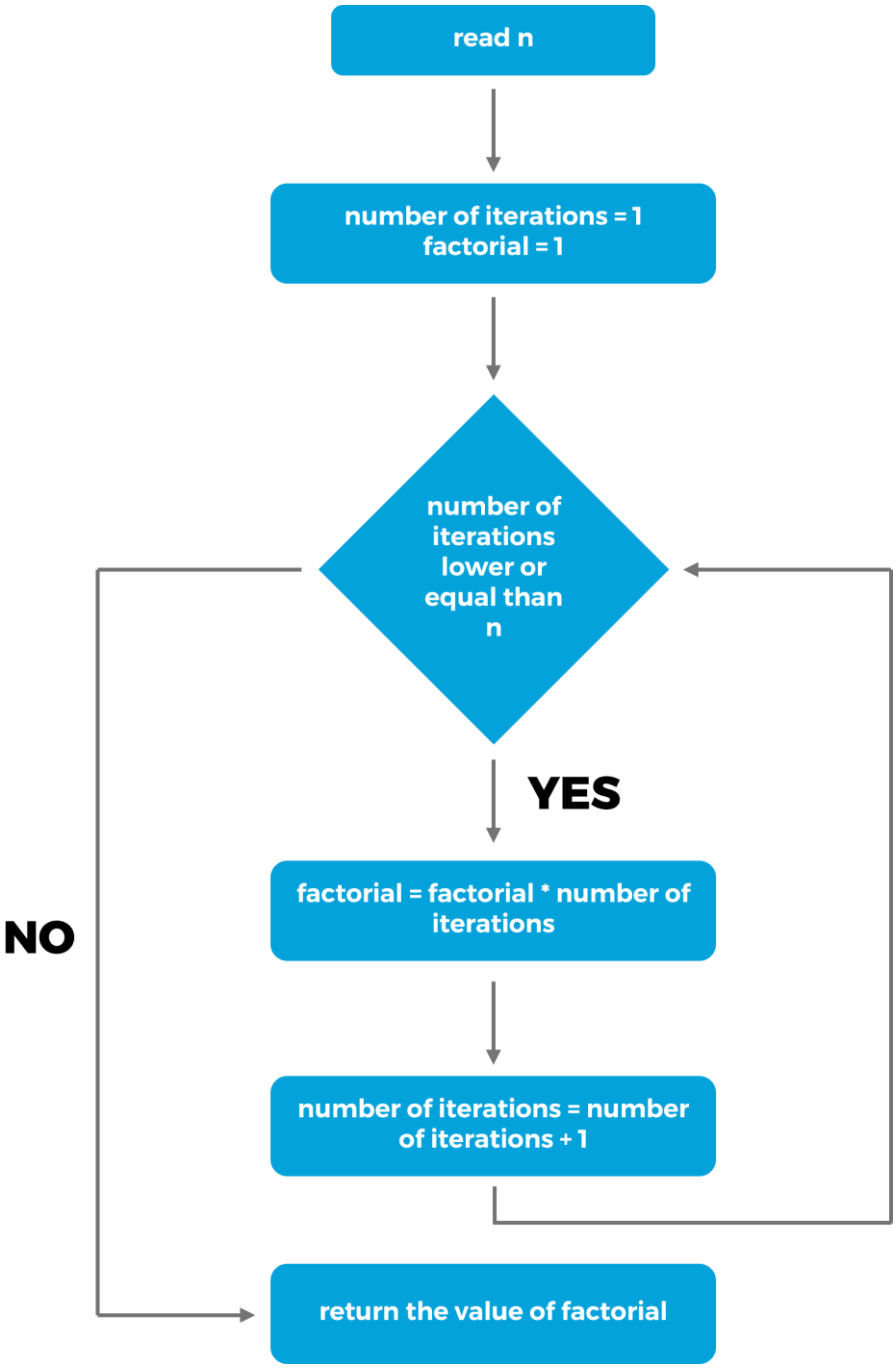
· **Repetitive or Iterative:** Repetition is the heart of these algorithms. They execute a set of instructions repeatedly until a specific condition is met. Loops (such as "for" and "while") fall into this category. They're essential for tasks like processing large datasets, simulating events, or solving optimization problems.

Example: A program that calculates the factorial of a number using a loop. The factorial of a number is a mathematical operation that multiplies a given number, n, by every integer less than n down to 1.

1. Read n, that is the number for which we want to calculate the factorial.
2. Set the value factorial is 1.
3. Loop through from 1 to n:
4. Multiply factorial by the value of the loop.
5. Show the value of factorial.

Note that the indentation of instruction number 4 also has a specific role in the algorithm definition. This instruction is executed n times once at each iteration.

Graphic 1. Loop example



Source | Own elaboration

We can construct more and more complex algorithms by combining these three techniques. Let's see an example of an algorithm that uses all composition schemes to find the maximum value in an integer in a list:

List with 10 integers. Number 10 is the first element of the list.

1. Current item is the first element of the list.
2. Maximum value is the current item.
3. While we don't evaluate all elements in the list, loop:
4. If the current item is greater than the maximum value then:
5. Maximum value is the current item.
6. Current item is the next element of the list.
7. Return the maximum value.

2.4. Algorithm Efficiency

In the field of computer science, algorithms are represented as programs. A program can be defined as the description of an algorithm in the finite repertoire of machine instructions. The set of programs that a computer equipment has is called software. To get a computer to carry out a task, we first must think and design an algorithm to solve it, then we must program it into the computer. The goal is to transform the conceptual algorithm into a clear set of instructions in a specific programming language, based on different approaches to the programming process (programming paradigms).

As problems become more complex, so do the algorithms that solve them. It is at this point that it becomes necessary to be able to evaluate the cost that an algorithm has. Algorithmic efficiency refers to how efficiently an algorithm utilizes computational resources. This property is essential for understanding an algorithm's performance and involves analysing its resource consumption, including time and space. Achieving maximum efficiency entails minimizing resource usage. However, comparing algorithms' efficiency can be complex because different resources, cannot be directly compared.

We can consider an algorithm efficient when its resource consumption, referred to as computational cost, remains at or below an acceptable threshold. This threshold is determined based on the algorithm's ability to operate within a reasonable timeframe or space on a standard computer, often relative to the input size. In essence, "acceptable" implies that the algorithm can execute within practical constraints, ensuring it remains viable for real-world usage.

There are two main measures of algorithmic efficiency:

Time Complexity: This is the computational complexity that describes the amount of time an algorithm takes to run as a function of the size of the input to the program.

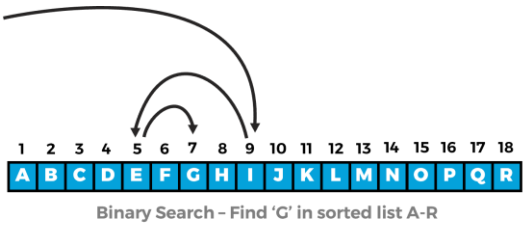
Space Complexity: This is the amount of memory that an algorithm needs to run to completion.

The goal is to minimize both time and space requirements. However, there's often a trade-off between these two. For example, an algorithm that runs faster might require more memory, and an algorithm that uses less memory might be slower. The choice of algorithm depends on the specific requirements of the task. Efficient algorithms strike a balance between these considerations, aiming for optimal performance in real-world scenarios.

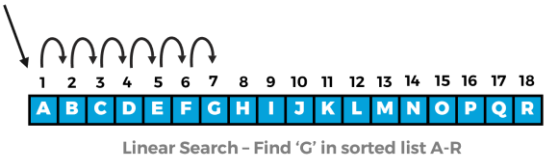
Let's see a simple example of two different algorithms that can be used to search an element in a list but have same space complexity but different time complexity. Then we will analyse when to use each one:

Linear search is an algorithm used to find a value within a list, it is very similar to the algorithm to find the maximum value of the list we have developed before. It evaluates each element of the list, starting from the beginning until the target value is found or the end of the list is reached. During each iteration, the algorithm compares the current element with the target value. If a match is found, the algorithm terminates, returning the index of the target value.

Binary search is an efficient algorithm used to find a target value within a sorted list by repeatedly dividing the search interval in half. It begins by examining the middle element of the list and compares it with the target value. If the middle element matches the target, the search is successful. Otherwise, if the target is less than the middle element, the search continues in the lower half of the list; if the target is greater, the search moves to the upper half. This process of halving the search interval is repeated until the target value is found or the search interval becomes empty. Binary search operates on the principle of divide and conquer, significantly reducing the search space with each iteration.



Binary Search – Find 'G' in sorted list A-R



Linear Search – Find 'G' in sorted list A-R

Source | Board Infinity

With these two examples we can easily understand that there exist multiple algorithms to perform the same task. Each algorithm has its own conditions, strengths, and weaknesses; Linear search is easy to implement and suitable for small lists and unsorted data. However, algorithm's efficiency decreases as the size of the list grows, making it less useful for large datasets as if the element we are searching is near the end of the end of the list we must evaluate most of its elements. Binary search is highly efficient, making it ideal for searching large and sorted datasets. However, it requires the data to be sorted initially, which can be a prerequisite for its application.

2.5. Limitations

Over the past few years, we've observed a remarkable surge in machine learning and artificial intelligence technologies, both of which heavily lean on algorithms to process insights from massive datasets. This surge has fuelled remarkable progress in various domains, including image and video recognition and generation, robotics, medical analysis, decision making, voice cloning or natural language processing. These advancements illustrate the boundless potential that algorithms offer, propelling us into an era marked by unprecedented innovation, but it is important to understand that even with these important advances, algorithms have its own limitations. Thera are problems that cannot be solved by an algorithm and there are problems that we are not prepared to solve.

First, we are going to describe the problems that an algorithm cannot solve. There are two categories: Undecidable problems are those that are theoretically impossible to solve by any algorithm and intractable problems, those problems that have no reasonable time solutions.

A classic example of undecidable problem is the halting problem, a decision problem with a binary (yes or no) answer that is undecidable. It poses the challenge of determining whether a computer program will eventually halt or run indefinitely. The statement of this problem is like: 'Is it possible to write a program that can tell given any program and its inputs and without executing the program, whether it will halt?'

On the other hand, there are the intractable problems. Some intractable problems are solvable in a reasonable time for small inputs only, but the algorithm doesn't scale to larger inputs. Some intractable problems are solved everyday even though they are intractable. Such has routing algorithms, timetabling, resource scheduling. These tend to use a guessed solution to an intractable problem which can then be checked quickly. This is called heuristics, it uses knowledge and experience to make intelligent guesses, the heuristic approach may also seek to find a suitable solution and not necessarily the perfect or ideal.

One of the most famous intractable problems is the **Travelling Salesman Problem** (TSP). In this challenge, the task is to find the shortest route that allows a salesman to visit each city on a map exactly once before returning to the starting city. The goal is to minimize the total distance travelled, presenting a great computational challenge due to the exponential increase in possibilities as the number of cities grows.

Let's make some numbers to observe how it grows. To calculate the number of roads between cities we must calculate the factorial of the number of cities minus one. The number of routes we must explore is the half of the number of connections between cities (we only must use each road once).

By looking at the table, we can observe how the number of possibilities grows very fast[3]:

Number of cities	Number of connections between cities	Number of possible routes
5	4! = 24	12
10	9! = 362.880	181.440
15	14! = 87.178.291.200	43.589.145.600

2.6. Limitations of machine learning algorithms

In the era of machine (deep) learning, we have a huge collection of algorithms designed to solve multiple tasks, for example: Decision trees, Random Forest, Support Vector Machines or Neural Networks (in all its versions) ... All of them can be adapted to different problems by a training process where we show them data and they adapt themselves. By this process we can obtain suitable solutions. It is important to note that them also have their own limitations.

- **Data Quality and Quantity:** Machine learning algorithms heavily rely on the quality and quantity of data available for training. Limited or biased data can lead to inaccurate or biased model predictions. Additionally, insufficient data may hinder the algorithm's ability to generalize well to unseen data.

- **Overfitting and Underfitting:** Machine learning models may suffer from overfitting, where they learn to capture noise or irrelevant patterns in the training data, leading to poor performance on new data. On the other hand, underfitting occurs when the model is too simplistic, resulting in suboptimal performance.

- **Interpretability:** Many machine learning models, especially complex ones like neural networks, lack interpretability, making it difficult to understand and

[3] Example taken from https://runestone.academy/ns/books/published/mobilecsp/Unit5-Algorithms-Procedural-Abstraction/Limits-of-Algorithms.html

trust their predictions. This can be problematic in domains where interpretability is crucial, such as healthcare and finance.

- **Computational Resources:** Training complex machine learning models often require significant computational resources, including high-performance hardware and large-scale data processing infrastructure. Limited computational resources can constrain the scalability and practicality of machine learning solutions.

- **Domain-specific Expertise:** Developing effective machine learning solutions often requires domain-specific expertise to properly preprocess data, engineer relevant features, and interpret model outputs.

- **Continuous Learning and Adaptation:** Machine learning algorithms may struggle to adapt to changing environments or evolving data distributions over time. Continuous monitoring and retraining of models are necessary to ensure their performance remains robust and up to date.

TIPS AND RECOMMENDATIONS FOR TEACHERS

Tips and recommendations for teachers to engage with algorithm complexity concepts

1. Build a Strong Foundation
Start with a basic overview of what algorithms are and their historical development to provide context. This helps build a narrative that makes the abstract concept more tangible and relatable.

2. Use Analogies and Metaphors
Compare algorithms to everyday activities like cooking or navigating a city. This can help demystify the concept and make it accessible to learners who may not have a strong technical background.

3. Interactive Demonstrations
Show algorithms in action through live coding sessions or interactive web tools that allow participants to input examples and see how algorithms process them. This can be particularly engaging and informative.

4. Foster an Interactive Classroom
Encourage questions and discussions to allow participants to express their understanding or confusion about the topics being covered. This also helps to gauge their engagement and comprehension in real-time.

5. Incremental Learning
Break down complex topics into smaller, manageable parts. Start with basic concepts and gradually introduce more complexity. This scaffolding approach helps prevent cognitive overload.

Warm-up Activities

1. Simple Algorithm Creation
Ask participants to write a simple algorithm for an everyday task, like making tea or getting ready for work. This activity gets them thinking in the step-by-step logical sequences that are fundamental to algorithm design.

2. Historical Algorithm Exploration
Discuss the Euclidean algorithm as a group, exploring its steps and how it was historically used to solve problems. This can lead to a deeper appreciation of how fundamental concepts in algorithms have long-standing roots.

3. Control Structure Role Play
Organize a role-play where each participant acts out a part of a control structure in an algorithm, such as a conditional or loop. This will help them visualize and understand how different parts of an algorithm interact.

4. Algorithm Debugging Challenge
Provide a simple algorithm with intentional errors and ask participants to 'debug' it. This can be a fun and engaging way to deepen their understanding of how algorithms are structured and function.

Evaluation Tips and Recommendations

1. Continuous Feedback
Provide immediate and ongoing feedback throughout the course. This helps participants correct misunderstandings early and reinforces their learning as they progress.

2. Practical Assignments
Design assignments that require participants to apply what they've learned in practical scenarios. For instance, they could optimize an existing algorithm or identify inefficiencies in a given algorithm.

3. Peer Review
Incorporate peer review sessions where participants evaluate each other's work. This not only provides multiple feedback points but also encourages collaborative learning and critical thinking.

4. Use Rubrics
Develop clear, criteria-based rubrics for evaluating assignments. This ensures consistency in grading and clear expectations for participants.

5. Self-Assessment
Encourage participants to assess their own work based on given criteria. This promotes self-reflection and deeper learning.

These strategies will help trainers effectively deliver complex content on algorithms while ensuring that participants are actively engaged and able to practically apply their knowledge.

3. Recognize the role of algorithms in various sectors and the potential pitfalls arising from their use

Gaining a deep understanding of **the role of algorithms across various sectors**, while simultaneously recognizing potential pitfalls, is integral to fostering a comprehensive and holistic view of the implications and responsibilities tied to the deployment of algorithms. This notion emphasizes the necessity for thorough insight into how algorithms can influence different aspects of society, economy, and governance. The diverse applications of algorithms, from enhancing operational efficiencies to personalizing user experiences, demonstrate their significant impact on modern life (Mourtzis et al., 2022). However, this influence also carries substantial risks, such as the potential for **bias**, **privacy** breaches, and the amplification of **existing societal inequalities** (Patel, 2024).

The deployment of algorithms must be guided by ethical considerations and **regulatory frameworks** to mitigate these risks (Tsamados et al., 2021). It is crucial to develop robust systems of accountability that ensure algorithms do not perpetuate or exacerbate harmful practices or disparities. This includes rigorous testing and validation processes that assess algorithms for **fairness** and accuracy, especially in sensitive areas such as healthcare, criminal justice, and financial services. Stakeholders, including technologists, policymakers, and the public, must engage in ongoing dialogues to examine the implications of algorithmic decisions and ensure these tools serve the broader interests of society (Donovan et al., 2018).

Moreover, **transparency** in algorithmic operations and decision-making processes is essential. The public needs assurances that algorithmic decisions are made with ethical considerations and are open to scrutiny. This transparency is crucial not only for building trust but also for enabling experts and regulators to perform audits and assessments effectively (Felzmann et al., 2019). Educational initiatives that improve algorithmic literacy among the public and policymakers are also vital (Reisdorf & Blank, 2021). Such efforts will empower more individuals to participate in discussions and make informed decisions about the use and regulation of these technologies.

Research and development in the field of algorithmic fairness and ethics must continue to evolve. As technology advances, new challenges will emerge, requiring innovative solutions that address both technical and ethical issues. Collaboration among academia, industry, and government can facilitate the sharing of best practices and promote the development of more equitable and responsible algorithms.

The conversation about algorithms also needs to include the potential for positive change. When designed and implemented responsibly, algorithms have the power to enhance accessibility, efficiency, and fairness. For example, in education, algorithms can help create personalized learning experiences that adapt to the individual needs of students, potentially closing gaps in educational attainment.

In healthcare, predictive algorithms can improve diagnostics and patient outcomes by identifying risks and recommending personalized treatment plans.

In conclusion, the integration of algorithms into various sectors presents both opportunities and challenges (Dwivedi et al., 2021). By fostering an environment that encourages ethical practices, transparency, and inclusivity, society can harness the benefits of algorithms while minimizing their risks. As Williamson (2018) suggests, a nuanced understanding of these dynamics is essential for realizing the full potential of algorithmic applications in a manner that is beneficial and just for all members of society.

3.1. Role of algorithms in various sectors

As highlighted by Parker et al. (2016), "algorithms have permeated every sector, revolutionizing processes and decision-making". In this section, we explore some prominent sectors where algorithmic influence is notably significant:

a. Finance: AI in finance has evolved significantly from its rudimentary beginnings to becoming an indispensable tool in the industry. The use of AI in finance can be traced back to the 1980s when basic programs were adapted to automate simple tasks such as data entry and accounting. However, the real transformative phase began in the late 1990s and early 2000s with the advent of more sophisticated machine learning algorithms and the explosion in data availability (Aksoy & Gurol, 2021). Financial institutions quickly recognized the potential of AI to provide more accurate risk assessments, faster transaction processing, and enhanced customer service. This was further fuelled by the increase in computational power and the development of advanced data analytics techniques. By the 2010s, AI had begun reshaping trading, underwriting, fraud detection, and customer relationship management in profound ways (Davenport & Mittal, 2023). Here are some examples of AI use in the financial sector:

- Algorithmic Trading: One of the most famous examples of successful algorithmic trading is Renaissance Technologies, a hedge fund known for its Medallion Fund. This fund uses complex mathematical models to analyse and execute trades at high speeds. By employing advanced algorithms, Renaissance has been able to achieve outstanding returns, far surpassing the performance of the market (Qin, 2012).
- Risk Management: JPMorgan Chase uses algorithms to manage credit risk. Their system assesses the risk of loan defaults based on numerous factors including economic conditions, sector performance, and individual credit history. This algorithmic approach allows them to tailor their loan offerings and minimize potential bad debt losses.
- Fraud Detection: PayPal employs machine learning algorithms to detect fraudulent transactions. These algorithms analyze thousands of transaction

attributes in real-time, including the amount, the device used, and the transaction history of the user. By identifying patterns that deviate from the norm, PayPal's systems can flag potentially fraudulent transactions with high accuracy, thereby preventing financial loss and protecting users.

- Customer Service: Bank of America uses an AI-driven virtual assistant named Erica. This algorithmic tool helps customers manage their accounts, track spending, and make payments. Erica can also provide real-time updates on account changes and suggest ways to save money based on spending patterns analysed by algorithms.

b. Healthcare: AI in healthcare has experienced a remarkable evolution, from early experimental applications to becoming a critical component of modern medical practice. This transformation has largely been fuelled by advancements in machine learning, big data analytics, and the increasing digitization of healthcare records. As we delve deeper, we will explore the various facets of AI implementation in healthcare, including diagnostic assistance, personalized treatment, disease prediction, patient management, and robotic surgeries (Bohr & Memarzadeh, 2020). Here are some remarkable examples:

- **Diagnostic Assistance:** AI's ability to analyse large datasets has revolutionized diagnostic processes in healthcare. By integrating AI with imaging technologies, medical professionals can detect diseases with higher accuracy and speed than traditional methods. IBM Watson Health demonstrates the power of AI in enhancing diagnostic accuracy. Watson's advanced AI algorithms can analyse the meaning and context of structured and unstructured data in clinical notes and reports. For instance, it has been used in oncology to identify treatment options for cancer patients by cross-referencing millions of oncology clinical notes with patient medical records and existing literature.
- **Personalized Treatment:** AI facilitates personalized medicine by considering individual genetic profiles, environmental factors, and lifestyle choices to tailor treatments. This approach significantly improves patient outcomes by targeting therapies that are most likely to work for specific patients. **Tempus** uses AI to gather and analyse vast amounts of genomic and clinical data to help doctors create personalized treatment plans for cancer patients. Its platform utilizes machine learning algorithms to understand molecular and therapeutic data and predict which treatments are likely to be most effective for individual patients.
- **Disease Prediction and Management**: AI models are increasingly used to predict the likelihood of disease development, progression, and potential complications. This proactive approach allows for earlier interventions, potentially saving lives and reducing healthcare costs. Google's DeepMind has developed AI systems that can predict acute kidney injury up to 48 hours before it happens with remarkable accuracy. The AI model analyses a wide range of health data in real time, enabling timely interventions that can prevent deterioration and improve outcomes.

- **Patient Management and Monitoring**: AI enhances patient management by continuously monitoring health data through wearable technology and IoT devices. These tools alert healthcare providers to changes in a patient's condition, allowing for immediate response and continuous care adjustments. Virta Health employs AI-driven tools to manage chronic diseases such as type 2 diabetes. Its platform monitors patient data in real-time and adjusts treatment plans dynamically, helping maintain optimal blood glucose levels and reducing dependency on medications.
- **Robotic Surgery**: Robotic surgery, assisted by AI, offers high precision, reduced trauma, and faster recovery times. AI algorithms provide real-time data to surgeons and can even automate certain aspects of surgery for improved safety and outcomes. Intuitive Surgical' s da Vinci robotic surgical system uses AI to enhance surgical precision. The system provides surgeons with highly magnified, 3D high-definition views of the surgical site and translates the surgeon's hand movements into smaller, more precise movements of tiny instruments inside the patient's body.

As AI continues to advance, its potential to transform healthcare grows exponentially. These examples underscore AI's ability to enhance the accuracy of diagnoses, tailor treatments to individual patients, predict disease courses, manage chronic conditions effectively, and execute surgical procedures with unprecedented precision. The integration of AI in healthcare not only optimizes patient outcomes but also streamlines the work of healthcare providers, paving the way for more innovative approaches to medicine and surgery. The ongoing development and deployment of AI technologies hold great promise for addressing some of the most challenging issues in healthcare today.

c. E-commerce: AI significantly improves both customer and **business experiences**, particularly in retail and e-commerce sectors. Through the accumulation of extensive consumer data and its integration into machine learning algorithms, retailers can develop advanced personalization, recommendation, and automation functionalities. These AI-driven features are now standard across shopping platforms and serve to enhance customer engagement and streamline operations, thereby boosting company profits and resource efficiency. These are some examples of AI use in e-commerce and retail businesses:

- **Enhanced Advertising Strategies**
  Smartly.io: Utilizes AI to automate and optimize social media advertising campaigns, significantly reducing manual effort while improving ad performance and engagement across platforms.
  eBay: Employs AI to offer personalized shopping recommendations and customer advice, boosting buyer satisfaction and retention.
- **Automotive Sales Optimization**
  Cox Automotive: Implements AI through its Esntial tool to streamline car buying processes online, including payment estimation and risk assessment, enhancing the digital customer journey.
- **Efficient Delivery Systems**

Gopuff: Integrates AI to optimize delivery routes, ensuring quick and efficient delivery of everyday essentials through its micro-fulfilment centres.

- **Innovative Product Development**
  Mondelez International: Leverages AI in its research and development sector to accelerate innovation and improve cost-efficiency in snack food production.
- **Personalized Shopping Experiences**
  Mirakl: Uses AI to tailor product recommendations, improving customer engagement and sales in digital marketplaces.
  Rue Gilt Groupe: Applies AI to refine product recommendations, enhancing the shopping experience on its luxury merchandise websites.
- **Advanced Inventory and Supply Chain Management**
  IBM Watson: Supports retail companies in streamlining operations and personalizing customer interactions through real-time data analysis.
  Anaplan: Utilizes predictive intelligence to forecast business outcomes and optimize retail strategies, impacting everything from inventory management to customer acquisition.
- **Counterfeit Detection and Content Moderation**
  3PM Solutions: Employs AI to detect counterfeit products and ensure the accuracy of seller ratings on major online marketplaces.
- **Conversational AI and Customer Engagement**
  Valyant AI: Develops AI-driven conversational tools for quick service restaurants, enhancing customer ordering experiences through voice-based technologies.

d. Transportation: There is no doubt, AI is reshaping the transportation industry by enhancing efficiency, safety, and user experiences. This section delves into the various ways AI is being integrated into transportation systems, from public transit and logistics to personal mobility and traffic management.

- **Autonomous Vehicles**
  Self-Driving Cars: AI powers the core functionalities of autonomous vehicles, enabling them to navigate, avoid obstacles, and make real-time decisions. Companies like Tesla and Waymo lead advancements in this area, significantly improving road safety and vehicle efficiency.
  Drones for Delivery: Companies like Amazon are experimenting with drones that use AI to deliver packages autonomously, promising to reduce delivery times and costs for last-mile logistics.
- **Traffic Management and Smart Cities**
  Intelligent Traffic Systems: AI is used to analyse traffic patterns and optimize traffic light sequences, reducing congestion and enhancing road safety. Cities like Barcelona and Singapore utilize AI systems to maintain smooth traffic flow across busy intersections.
  Predictive Maintenance: AI helps predict when public transport vehicles and infrastructure (like bridges and roads) will need maintenance, preventing breakdowns and extending the lifespan of public assets.

- **Public Transit Optimization**
  Route Planning: AI algorithms optimize bus and train schedules based on real-time data and historical traffic patterns, maximizing operational efficiency and minimizing wait times for passengers.
  Capacity Management: During peak times, **AI systems can predict passenger flow and adjust the frequency of transit services accordingly, enhancing commuter convenience.**
- **Freight and Logistics**
  Automated Warehousing: Companies like Amazon use AI-driven robots to streamline warehouse operations, increasing the speed and accuracy of picking and packing processes.
  Optimized Routing: AI tools analyse numerous variables such as weather, traffic conditions, and delivery windows to suggest the most efficient routes for shipment, saving time and fuel costs.
- **Enhanced Safety Features**
  Collision Avoidance Systems: AI-powered sensors and onboard systems in vehicles can identify potential hazards and take immediate action to avoid accidents.
  Driver Monitoring Systems: These systems use AI to monitor drivers' alertness and overall health to prevent accidents caused by driver fatigue or medical emergencies.
- **Customer Service and Support**
  AI Chatbots: Transportation companies employ AI chatbots to provide real-time assistance to travellers, handling inquiries about schedules, fares, and disruptions without human intervention.
  Personalized Travel Recommendations: AI algorithms analyse user preferences and past behaviour to offer customized travel suggestions, enhancing the customer experience.

e. Education: AI is transforming the educational landscape by enabling personalized learning experiences, automating administrative tasks, and facilitating immersive educational environments. This section explores the diverse applications of AI in education, highlighting how it supports teachers, enhances student learning, and optimizes educational management.

- **Personalized Learning**
  Adaptive Learning Platforms: AI systems like DreamBox Learning and Knewton provide personalized learning experiences by adapting to individual student's pace and learning style. These platforms assess students' knowledge levels and automatically adjust content to suit their learning needs.
  AI Tutors and Assistants: AI-driven tutoring systems such as Carnegie Learning offer real-time feedback and assistance to students, helping them understand complex subjects through personalized instruction and practice.
- **Educational Content Development**
  Automated Content Generation: AI tools can help in creating and customizing educational content. Tools like Quillionz use AI to generate

quizzes and summaries from textbooks and articles, enhancing study materials without extensive human input.
Language Learning Apps: Applications like Duolingo use AI to tailor language learning courses to the user's proficiency and progress, making language learning more accessible and effective.

- **Assessment and Feedback**
  Automated Grading Systems: AI automates the grading of standardized tests and even essay writing, reducing the administrative burden on educators and allowing them to focus more on teaching and less on grading.
  Predictive Analytics: AI systems analyse student data to predict academic risks and outcomes, enabling educators to intervene early with at-risk students to improve their learning outcomes.
- **Administrative Automation**
  Enrolment and Admissions Processes: AI streamlines administrative tasks such as admissions and enrolment by automating data processing and decision-making, thus reducing paperwork and improving accuracy.
  Resource Management: AI tools help manage school resources, scheduling, and student attendance, optimizing operations and reducing overhead costs.
- **Enhanced Accessibility**
  Assistive Technologies: AI-powered tools like speech recognition and text-to-speech converters make learning materials more accessible to students with disabilities, ensuring inclusivity and equal opportunities for all learners.
  Visual and Auditory Learning Aids: AI-driven applications such as Microsoft's Immersive Reader help students with dyslexia by offering reading assistance, proving that technology can bridge learning gaps.
- **Immersive Learning Environments**
  Virtual Reality (VR) and Augmented Reality (AR): AI integrated with VR and AR creates immersive learning experiences that simulate real-world scenarios, making complex subjects like science and history more engaging and understandable.
  Educational Games and Simulations: AI-enhanced educational games adapt to student responses, providing a dynamic learning environment that stimulates engagement and enhances learning.

f. Security and military sector: AI is increasingly pivotal in enhancing security measures and military operations. This section outlines how AI technologies are integrated into defence strategies, surveillance systems, and security protocols, improving efficiency and effectiveness while addressing complex challenges in national and global security.

- **Enhanced Surveillance and Monitoring**
  Automated Surveillance Systems: AI-driven systems are employed to monitor sensitive areas, using real-time data analysis to detect unusual activities or threats. These systems can differentiate between normal and

suspicious behaviours, significantly reducing human error and response times.
Drone Surveillance: AI-powered drones are used for aerial surveillance, providing comprehensive reconnaissance over hard-to-reach areas or conflict zones without risking human lives.

- **Cybersecurity and Défense**
  Threat Detection and Response: AI systems analyse patterns in network traffic to identify potential cyber threats, such as malware and ransomware attacks, much faster than human operators could. Once threats are detected, AI can also automate responses or recommend mitigation strategies.
  Data Encryption: AI enhances encryption techniques, making it more difficult for unauthorized entities to decipher sensitive information, thus securing data transmissions in military and security networks.
- **Autonomous Défense Systems**
  Robotic Combat Units: AI is used to operate autonomous or semi-autonomous robotic units that can perform various military tasks, from reconnaissance to combat roles, reducing the need for human soldiers in dangerous operations.
  Missile Défense Systems: AI algorithms help predict and intercept incoming threats with high precision, such as missiles or unmanned aerial vehicles (UAVs), ensuring proactive defence postures.
- **Intelligence Analysis and Decision Support**
  Data Integration and Analysis: AI systems integrate and analyse vast amounts of intelligence from multiple sources, providing a comprehensive overview that helps military strategists make informed decisions quickly and accurately.
  Simulation and Training: AI-driven simulations and virtual reality (VR) environments provide realistic training for security personnel and soldiers, preparing them for a variety of scenarios without the risks of live combat.
- **Maintenance and Logistics**
  Predictive Maintenance: AI tools predict when military equipment needs maintenance, which helps prevent malfunctions and extends the lifespan of valuable assets.
  Supply Chain Optimization: AI optimizes logistics and supply chain management, ensuring that resources are efficiently allocated and delivered where needed, especially in complex military operations.

g. Media and leisure: AI has become a transformative force in the media and leisure industries, enhancing content creation, personalizing user experiences, and optimizing operations. This section explores how AI technologies are being integrated into various aspects of media production, distribution, and leisure activities.

- **Content Creation and Management**
  Automated Journalism: AI tools such as natural language generation (NLG) software are used to create news articles and reports, especially for data-driven content like sports results and financial updates, allowing for

quicker publishing and freeing up human journalists for in-depth investigations.
Video and Music Production: AI algorithms assist in editing, from colour correction in videos to sound mixing in music, streamlining production processes and enabling more creative experimentation.

- **Personalization and Recommendation Systems**
  Media Recommendations: Streaming platforms like Netflix and Spotify use AI to analyse viewing and listening habits, respectively, offering personalized content recommendations to keep users engaged and improve satisfaction.
  Targeted Advertising: AI-driven analysis of user data helps media companies create and deliver targeted advertisements, increasing efficiency and effectiveness in marketing campaigns.
- **Audience Engagement and Interactive Media**
  Chatbots and Virtual Assistants: AI-powered chatbots provide real-time interaction with users, offering customer support, content recommendations, and interactive experiences in gaming and virtual reality (VR) environments.
  Augmented Reality (AR) and VR: AI is integral to developing immersive AR and VR experiences, enhancing the realism and interactivity of virtual environments used in games, virtual tours, and educational content.
- **Operational Efficiency**
  Programmatic Advertising: AI automates the buying and selling of ad space, optimizing placement and pricing in real time to maximize return on investment for advertisers and media outlets.
  Content Distribution: AI tools help optimize distribution strategies, analysing consumption patterns to determine the best channels and times for releasing content to maximize reach and engagement.
- **Enhancing Creative Processes**
  Scriptwriting and Plot Development: AI tools aid in scriptwriting by suggesting plot developments, dialogue, and character interactions based on genre conventions and audience preferences.
  Art and Graphics Design: AI-driven tools assist artists and designers by suggesting improvements, generating ideas, and automating repetitive tasks, allowing for greater creative freedom and experimentation.
- **Event Management and Planning**
  Crowd Management: AI is used in planning and managing large events, analysing attendee data to optimize venue layout, scheduling, and security measures.
  Personalized Experiences: At leisure parks and events, AI systems offer personalized itineraries and experiences based on visitor preferences and past behaviour, enhancing customer satisfaction and retention.

3.2. Potential pitfalls of algorithm use

Despite their numerous advantages, algorithms can precipitate significant pitfalls such as bias and discrimination, privacy issues, lack of transparency and accountability, and an overreliance on automation (O'Neil, 2016). Each of these pitfalls not only poses ethical and operational risks but also could undermine public trust in AI technologies. Here, we explore these categories in more detail, providing examples from various sectors to illustrate the potential dangers and missteps in AI implementation.

a. Bias and Discrimination: AI systems can inadvertently perpetuate and amplify biases if they are trained on datasets that are not representative or contain prejudicial errors. This issue is particularly significant in sectors like finance and healthcare, where such biases can lead to unfair treatment of individuals.

- Finance: AI systems used in credit scoring can reflect existing racial or gender biases present in historical data. For instance, if a model is trained on past loan approval data, and those data reflect historical biases against a particular demographic group, the model may also discriminate against that group.
- Healthcare: There have been instances where AI diagnostic tools have shown discrepancies in treatment recommendations based on racial or gender differences. A notable example is an AI system that misdiagnosed certain skin diseases in darker-skinned individuals due to a dataset predominantly consisting of lighter skin tones.

b. Privacy Concerns: The extensive data required to train and operate AI systems raise significant privacy concerns, particularly about how data is collected, stored, and used. This is evident in sectors like e-commerce and education, where personal data is extensively used to enhance customer and student experiences.

- E-commerce: Retailers using AI to personalize shopping experiences may collect vast amounts of personal information, ranging from purchase history to real-time location data. This raises concerns about the potential for data breaches and the unauthorized use of personal information.
- Education: In the educational sector, AI systems used to track student progress can collect sensitive information about learning difficulties, health data, and even behavioural patterns. This poses risks regarding who has access to this data and how it might be used beyond the educational context.

c. Lack of Transparency and Accountability: AI systems often operate as **"black boxes"**, with decision-making processes that are obscure or hidden from users and regulators. This lack of transparency can lead to accountability issues, especially when decisions have significant consequences on people's lives.

- Security and Military: In military applications, AI-driven decision-making in autonomous weapons can lead to life-or-death outcomes. The lack of transparency in how these decisions are made complicates the ethical implications and raises significant accountability concerns.
- Healthcare: An AI system used in patient diagnosis may not provide insight into its decision-making process, making it difficult for doctors to understand how it arrived at a certain diagnosis. This not only makes it harder to trust the AI's judgment but also complicates liability in cases of misdiagnosis.

d. Overreliance on Automation: Relying heavily on AI can lead to a degradation of human expertise, as skills atrophy when AI systems take over decision-making processes. This overreliance can be particularly problematic in high-stakes environments like transportation and finance.

- Transportation: The aviation industry has seen incidents where overreliance on automated systems has contributed to accidents. Pilots sometimes over-depend on autopilot and may lack sufficient manual flying experience to take over effectively in case of system failure.
- Finance: **The stock market has experienced several flash crashes, partly attributed to automated trading algorithms**. Overreliance on these systems without sufficient human oversight can lead to sudden market downturns based on algorithmic errors.

TIPS AND RECOMMENDATIONS FOR TEACHERS

Tips and recommendations for teachers to engage with the role of algorithms across various sectors

1. Use Case Studies
Start sessions by examining real-world case studies where algorithms have played a crucial role in various sectors. This provides a practical understanding of how abstract concepts are applied in different industries like finance, healthcare, and e-commerce.

2. Highlight Ethical Considerations
Emphasize the ethical implications of algorithm use, discussing potential biases, privacy concerns, and accountability. This will foster critical thinking and ethical awareness among participants.

3. Encourage Interactive Discussions
Facilitate group discussions on the implications of algorithm deployment. Pose questions like "What could go wrong with this algorithm in a real-world application?" or "How can we mitigate potential risks?"

4. Utilize Multimedia Resources

Incorporate videos, infographics, and podcasts that illustrate the impact of algorithms. This variety can cater to different learning styles and keep the content engaging.

5. Simulate Algorithm Impact
Use simulations or software tools that allow participants to see the effects of algorithm changes in simulated environments. This hands-on approach helps solidify understanding through practical application.

Warm-up activities

1. Algorithm Role Mapping
Assign roles to participants where each person represents a component of an algorithm in a specific sector. They must explain their role's impact and potential pitfalls, promoting an immersive understanding of algorithm functions and ethical implications.

2. Debate on Algorithm Ethics
Organize a debate on a controversial use of algorithms, such as facial recognition or predictive policing. This encourages participants to explore and articulate different perspectives on ethical challenges.

3. Scenario Analysis
Present small groups with scenarios involving algorithm use in various sectors, asking them to identify potential benefits, risks, and the ethical decisions makers would need to consider.

Evaluation tips and recommendations

1. Case Study Analysis
Assign participants to analyse a case study where algorithms are used in a specific sector. Assess their ability to identify both the benefits and the ethical considerations as described in the case.

2. Group Presentations
Have participants present on different aspects of algorithms' roles and risks in various sectors. Evaluate their understanding based on the depth of analysis and the ability to engage with ethical considerations.

3. Reflective Essays
Ask participants to write essays reflecting on how an understanding of algorithms can influence their professional responsibilities and ethical considerations. Assess their ability to integrate course content with personal or hypothetical professional scenarios.

4. Quizzes and Tests
Use quizzes to assess understanding of key concepts like algorithmic bias and transparency. Include scenario-based questions that require critical thinking beyond rote memorization.

5. Peer Feedback

Incorporate peer assessment as part of the evaluation process. This not only provides multiple perspectives on the participant's understanding but also encourages engagement with the learning material from a reviewer's standpoint.

Using these strategies, trainers can effectively educate participants on the significant impact of algorithms across sectors, ensuring they understand both the technological aspects and the broader ethical implications.

4. Comprehend algorithmic complexity, optimization, and the limitations of algorithmic decision-making

4.1. Algorithmic Complexity

In computer science, algorithms can either make or break the success of systems and apps because they need to run efficiently. Algorithmic complexity is like a guiding light for developers and engineers. It's all about figuring out how much computing power different algorithms need, and we use the **Big O notation** to do that. But it's not just about making things run faster; understanding algorithmic complexity helps us know how well our algorithms will work as we make our systems bigger, and how to manage the resources they need.

Algorithmic complexity is basically about how fast and how much space an algorithm needs to get a job done. Time complexity tells us how long it takes for an algorithm to solve a problem based on how big the problem is. Space complexity, on the other hand, is about how much memory the algorithm uses while it's working. These measurements help us figure out how well an algorithm works and if it can handle big tasks without slowing down too much.

In the world of computer science, there's this thing called Big O notation, which is the math way to talk about how fast or slow algorithms run. It helps us understand how much time or space an algorithm needs as we give it more stuff to work with. So, if we say an algorithm is O(n), it means that as we give it more input (let's call it "n"), it takes linearly more time to finish. But if it's O(n^2), then it'll take way more time as you increase 'n', because it's growing quadratically. So, Big O notation is like a quick way to understand how efficient an algorithm is going to be when it's dealing with a lot of data.

The importance of analysing algorithmic complexity spans various areas:

- **Predicting Performance:** When we analyse how complex algorithms are, we can anticipate how they will perform with different inputs. This ability to predict performance is crucial for choosing the most appropriate algorithm for a particular problem and for optimizing system performance.

- **Assessing Scalability:** In environments where input sizes can change, understanding algorithmic complexity helps us evaluate how well algorithms can

scale. Algorithms with favourable complexity characteristics perform reliably across a wide range of input sizes, ensuring scalability and responsiveness.

- **Managing Resources:** Efficient use of resources is critical in environments where resources are limited, such as embedded systems and cloud computing. Analysing algorithmic complexity informs decisions about how to allocate resources, ensuring optimal use and minimizing waste.

- **Designing Algorithms:** Understanding algorithmic complexity influences the design process, guiding developers toward algorithms that balance efficiency and functionality. By selecting algorithms with favourable complexity characteristics, developers can create systems that perform optimally without sacrificing functionality.

4..2. Time algorithmic complexity

Consider the ubiquitous problem of sorting. Algorithms such as bubble sort, which have an algorithmic complexity denoted as O(n2), tend to perform poorly when dealing with extensive datasets compared to more efficient alternatives like merge sort or quicksort, which boast an algorithmic complexity of O(n log n). Understanding this difference is pivotal for making well-informed decisions when selecting algorithms. This, in turn, leads to enhancements in system performance and subsequently enhances user experience (Karp, 2010).

For a clearer illustration, let's consider a scenario where we need to sort a dataset containing one billion entries (n = 1,000,000,000), which might be akin to sorting the users of a large social network.

- **Bubble Sort:** With an algorithmic complexity of O(n2), when applied to this massive dataset, the execution time would be astronomical. To give an estimate, the execution time would be proportional to approximately 1018, resulting in a staggering number. If each entry requires 1 nanosecond for sorting, the total time required by bubble sort to process one billion entries would amount to 1018 nanoseconds, equivalent to 11574.07 days or approximately 31.6881 years. This inefficient process would likely lead to significant delays in sorting the data, causing a poor user experience.

- **Quicksort:** On the other hand, with a time complexity of O(n log n), quicksort proves to be significantly more efficient. When sorting a dataset of one billion entries, the execution time using quicksort would be proportional to approximately 1,000,000,000 * log2(1,000,000,000). If each entry requires 1 nanosecond for sorting, the execution time required for quicksort to sort a dataset containing one billion entries would be approximately 29,897,352,853.986263 nanoseconds, equivalent to 29.8974 seconds, or approximately 0.00034603 days. Consequently, users would experience quicker response times and an overall smoother user experience.

Implementing more efficient sorting algorithms, such as quicksort, illustrates the transformative potential of optimization in computational processes (Bentley, 1999)

Why do we put the example of sorting? Organizing data is crucial for algorithms to work efficiently, akin to discovering a wealth of benefits such as faster processing and smoother user interactions. Imagine searching through a cluttered room for a specific item; it's time-consuming and laborious. Similarly, unsorted data in algorithms necessitates a linear search, where each item is inspected until the desired one is found. This process, with a complexity of O(n) (representing the dataset's size), resembles scanning through a book page by page to locate a word—a functional but inefficient method, especially for large datasets.

Now, picture the same room tidied up, items neatly arranged. Finding what you need becomes effortless because you know where to look. Similarly, sorted data enables algorithms to utilize binary search, a significantly more efficient approach. Binary search divides the search interval in half repeatedly until finding the desired item, with a time complexity of O(log n), scaling better with larger datasets. It's like locating a word in a dictionary by flipping through pages, halving the search space with each turn.

In essence, sorting data enables algorithms to work smarter, not harder. It converts a laborious linear search into a lightning-fast binary search, reducing processing times and enhancing overall performance. So, the next time you encounter unsorted data, remember the impact of sorting and its ability to unlock algorithmic efficiency.

4.3. Memory algorithmic complexity

When delving into algorithmic complexity concerning memory, one encounters a critical aspect of algorithm analysis that complements the examination of time complexity. Memory, a finite resource in computing environments, profoundly influences algorithm performance and system efficiency. Exploring the concept of algorithmic complexity related to memory can sound complex, but let's break it down into simpler terms.

When we talk about algorithmic complexity and memory, we're basically looking at how much space an algorithm needs to solve a problem. Think of it like organizing your bag for different tasks – some tasks might need more space, while others need less.

Imagine you're sorting a bunch of cards with numbers written on them. One way to do this is called merge sort. Merge sort divides the cards into smaller groups, sorts each group, and then puts them back together. While merge sort is really good at sorting quickly, it also needs some extra space to work with. This extra space is like having extra tables to organize your cards while you're sorting them.

Another example is when we're figuring out Fibonacci numbers, like those found in nature (like the pattern of petals on a flower or the spiral of a seashell). There's a way to calculate these numbers using a method called dynamic programming. This method helps us find the numbers faster, but it also needs some extra space to remember the numbers it's already calculated. It's like having a notebook where you jot down the numbers as you figure them out.

Now, let's talk about how we store information in computer programs. We have different tools called data structures, like arrays and lists, that help us keep things organized. Each of these tools takes up a certain amount of space in the computer's memory.

For example, imagine you're making a list of your friends' names. If you use a simple list, like writing their names on a piece of paper, it doesn't take up much space. But if you use something fancier, like a table with lots of columns, it might take up more space because you're storing extra information about each friend.

So, when we're analysing algorithmic complexity related to memory, we're basically thinking about how much space our algorithms and data structures need to do their jobs efficiently. By understanding this, we can make better choices in programming to make sure our programs run smoothly and don't use up too much memory. It's like organizing your bag smartly so you can carry everything you need without it getting too heavy.

In short, thinking about memory and algorithmic complexity helps us write better programs that work well and don't hog all the computer's resources. It's like being a smart organizer for your computer!

Algorithmic complexity is a crucial concept in modern computing. It offers a structured way to assess and enhance how efficient algorithms are. By using Big O notation, we can understand the basic traits of algorithms and how they perform with different amounts of data. Understanding algorithmic complexity helps developers and engineers create systems that can handle the constantly changing demands of computing. As technology progresses, knowing about algorithmic complexity will continue to be vital for driving innovation and improving what computers can do.

4.4. Optimisation

Optimization stands as a beacon of innovation and efficiency, guiding algorithms to perform with unparalleled speed and precision. At its core, optimization is the art of refining algorithms to minimize complexity and maximize performance, unleashing their full potential to tackle complex problems with finesse.

Imagine algorithms as the engines driving our digital world, powering everything from search engines to logistics systems. Just like a well-tuned engine, an optimized algorithm operates smoothly, swiftly navigating through vast datasets and intricate calculations. But how exactly does optimization achieve this feat?

Optimization encompasses a myriad of strategies aimed at streamlining algorithms. It involves identifying bottlenecks, eliminating redundant steps, and fine-tuning processes to ensure peak efficiency. At its essence, optimization is about finding the optimal balance between resource utilization and output quality.

One of the fundamental goals of optimization is to minimize complexity. Algorithms often encounter challenges when tasked with solving complex problems that require extensive computational resources. By optimizing algorithms, we aim to simplify their structure and reduce the computational burden, making them more manageable and scalable.

Moreover, optimization seeks to enhance performance by leveraging innovative approaches. This involves exploring cutting-edge techniques such as parallel computing, heuristic algorithms, and machine learning to boost efficiency. By embracing these advancements, optimization empowers algorithms to deliver faster results while consuming fewer resources. Consider, for instance, the optimization of search algorithms used by internet browsers. Through meticulous fine-tuning and algorithmic enhancements, search engines can swiftly sift through vast amounts of data to deliver relevant results in milliseconds. This optimization not only enhances user experience but also reduces server load and energy consumption.

Focusing on parallel computing, In the realm of computing, Graphics Processing Units (GPUs) have emerged as indispensable components, especially in the context of Artificial Intelligence (AI) and machine learning. Despite their widespread use, many are still unfamiliar with the inner workings of GPUs and their pivotal role in advancing AI technologies.

At its core, a GPU is a specialized electronic circuit designed to rapidly manipulate and alter memory to accelerate the creation of images in a frame buffer intended for output to a display device. Originally developed for rendering graphics, GPUs have evolved into highly parallelized processors capable of executing thousands of computational tasks simultaneously.

Unlike traditional Central Processing Units (CPUs), which excel at sequential processing tasks, GPUs are optimized for parallel processing. They consist of numerous smaller processing units called "cores," each capable of executing tasks independently. This parallel architecture enables GPUs to tackle computationally intensive tasks with remarkable efficiency. GPU was designed for parallelizing tasks that involve performing the same operation on multiple pieces of data simultaneously. This parallelism allows GPUs to handle computationally intensive tasks with remarkable efficiency (graphics rendering and image and video processing that involves performing numerous calculations for each pixel, simulation and modelling, cryptocurrency mining, and parallelized matrix operations, among others).

Parallelism is particularly advantageous in AI applications, where tasks often involve processing vast amounts of data simultaneously. Tasks such as training

deep neural networks, image recognition, natural language processing, and data analysis benefit immensely from the parallel processing prowess of GPUs.

AI relies heavily on complex mathematical operations, such as matrix multiplications and convolutions, which are fundamental to training and deploying machine learning models. These operations involve manipulating large matrices and performing numerous calculations concurrently, making them ideal candidates for parallel execution on GPUs.

At this point, we need to introduce power consumption of AI algorithms. As these systems have become ubiquitous across various domains, needing a focus on energy efficiency to mitigate environmental impact and operational costs. Optimization techniques and algorithmic complexity play pivotal roles in achieving energy savings within AI frameworks. Optimization involves refining algorithms for maximal task performance with minimal computational resources. Energy consumption in AI frameworks arises from tasks like data manipulation and model training. Low algorithmic complexity is crucial for reducing energy consumption as it requires fewer computational operations.

4.5. Limitations of algorithmic decision-making

Algorithmic decision-making has emerged as tool for automating and optimizing processes across diverse domains. However, while its promising, algorithmic decision-making harbours inherent limitations that warrant meticulous scrutiny. From finance to healthcare, education to criminal justice, algorithms have been deployed, promising unparalleled efficiency, accuracy, and objectivity. Anyway, there exists a complex network of constraints that hinder the unrestricted effectiveness of algorithmic decision-making (Yeung, 2018).

Bias and Discrimination: One of the predominant and detrimental constraints affecting algorithmic decision-making is the inclination towards bias and discrimination. Despite endeavours towards impartiality, algorithms frequently inherit biases present in the training data they utilize. The presence of historical injustices and societal biases within this data can sustain and intensify disparities, resulting in discriminatory results. Additionally, the opaqueness of algorithmic methodologies presents difficulties in identifying and mitigating biases, thereby heightening apprehensions regarding fairness and equality in decision-making processes. This can be a positive opportunity to detect biases and discrimination previously unnoticed by human observers. Through the analysis of large datasets using advanced algorithms, subtle patterns and disparities can be identified, enabling corrective actions to be taken to rectify systemic biases. The transparency and traceability features of certain algorithms allow for close scrutiny of the decision-making process, empowering stakeholders to address biases at their core.

Lack of Contextual Understanding: Algorithmic decision-making operates within a contextually naive domain, dependent solely on quantitative metrics and predefined rules. This constraint is especially pronounced in intricate and dynamic environments where contextual subtleties are integral to decision-making. Human judgment, augmented by contextual awareness and domain proficiency, frequently exceeds algorithmic capabilities in discerning nuanced intricacies and rendering informed decisions. As a result, algorithmic systems may manifest rigidity and insufficiency in accommodating evolving contexts, thereby compromising the effectiveness and applicability of their decisions.

Unforeseen Consequences and Ethical Dilemmas: The deterministic nature of algorithms renders them susceptible to unforeseen ramifications and ethical quandaries, stemming from the inherent oversimplification and reductionism in algorithmic decision-making processes. Unanticipated outcomes, cascading effects, and unintended repercussions may emerge due to algorithms' incapacity to accommodate the complexity and intricacy of real-world scenarios. Moreover, ethical dilemmas concerning privacy infringement, erosion of autonomy, and moral accountability underscore the necessity for meticulous deliberation and oversight in the implementation of algorithmic systems.

4.6. How to identify the potential sources of bias and errors of algorithmic systems? A crucial issue for ensuring the fairness, transparency, and accountability

One of the primary sources of bias in algorithmic decision-making stems from the data used to train and test the algorithms. Biases present in the data can propagate throughout the decision-making process, resulting in skewed outcomes. To identify potential biases in the data it is necessary to conduct a thorough analysis of the data collection process to assess for any systematic biases or underrepresentation of certain groups, to examine the demographic characteristics of the dataset to identify disparities or imbalances that may lead to biased outcomes, and to evaluate the data preprocessing steps, such as data cleaning and feature selection, to identify any unintended transformations or distortions of the data.

To identify potential biases and errors in the design and implementation of algorithms it should be scrutinized the algorithmic models and methodologies assessing the underlying assumptions and constraints of the algorithm to determine if they align with ethical and societal norms, examining the training process to identify potential sources of overfitting or underfitting, which may lead to inaccurate or unfair predictions, investigating the choice of features and variables used by the algorithm to ensure they are relevant, non-discriminatory, and representative of the population. Overfitting and underfitting refer to two common issues that arise when training machine learning models. Overfitting

occurs when a model learns to capture noise or random fluctuations in the training data, rather than the underlying patterns or relationships. As a result, the model performs well on the training data but fails to generalize to new, unseen data. On the other hand, underfitting occurs when a model is too simplistic to capture the underlying structure of the data, resulting in poor performance both on the training data and on new data.

To identify potential biases and errors in evaluation and validation it should be tested methodologies to measure the performance metrics of the algorithm across different demographic groups to detect disparities or inequities, to conduct sensitivity analyses to assess the robustness of the algorithm to variations in input data or model parameters, and to request feedback from stakeholders, including affected communities and domain experts, to identify potential biases or ethical concerns overlooked during development.

Also, it is necessary a continuous monitoring and auditing to detect and mitigate biases, errors, or unintended consequences.

4.7. Addressing and Mitigating Algorithmic Risks and Biases

Once identified, algorithmic risks and biases must be addressed through a multifaceted approach encompassing technical, procedural, and ethical considerations. Technical interventions may involve refining algorithms through techniques such as debiasing algorithms, diversifying training data, and incorporating fairness metrics into model evaluation. Procedural interventions entail implementing robust validation protocols, transparency measures, and accountability mechanisms to ensure algorithmic decisions are scrutinized and validated effectively. Ethical interventions revolve around fostering a culture of ethical awareness and responsibility among developers, policymakers, and end-users, emphasizing the ethical implications of algorithmic decision-making and promoting ethical guidelines and standards.

Mitigating algorithmic risks and biases requires ongoing monitoring, evaluation, and adaptation to evolving challenges and contexts. Continuous monitoring of algorithmic performance and impact enables stakeholders to detect and address emerging biases and risks in a timely manner. Evaluation mechanisms such as bias audits, sensitivity analyses, and impact assessments provide insights into the effectiveness of risk mitigation strategies and inform iterative improvements. Additionally, fostering interdisciplinary collaboration and engagement facilitates the exchange of insights and best practices across domains, enriching the collective effort to minimize algorithmic risks and biases.

Tips and recommendations for teachers to engage with algorithmic complexity

1. Conceptual Introduction
Begin with a straightforward explanation of what algorithmic complexity is, using analogies related to everyday activities that require planning and resources, such as organizing an event or planning a trip, to illustrate concepts of efficiency and resource management.

2. Visual Demonstrations
Use visual aids, such as charts and graphs, to demonstrate how different algorithms perform under various conditions. This helps participants visually understand how the complexity affects performance.

3. Hands-on Activities
Incorporate practical exercises where participants can experiment with different algorithms to solve specific problems. This could involve simple coding exercises or using software tools that simulate algorithm performance.

4. Discuss Real-world Applications
Link the discussion to real-world applications where algorithmic complexity plays a critical role, such as data processing in large tech companies or route finding in GPS navigation systems, to highlight its importance.

5. Encourage Collaborative Learning
Facilitate group discussions and problem-solving sessions where learners can debate the best algorithms to use in different scenarios, promoting deeper understanding through collaboration.

Warm-up activities

1. Sorting Challenge
Organize a hands-on activity where participants manually sort objects (like books or cards) using different methods to mimic sorting algorithms. Discuss how each method relates to an algorithm's time and space complexity.

2. Algorithm Complexity Matching Game
Create a card game where participants match different algorithms with their Big O notations and examples of optimal use cases. This interactive approach helps solidify understanding of theoretical concepts through play.

3. Efficiency Brainstorm
Have participants list out daily tasks they perform that could be optimized with algorithms (like sorting emails or organizing files). Discuss how understanding algorithmic complexity can lead to improvements in these tasks.

Evaluation tips and recommendations

1. Concept Checks
Use frequent short quizzes to assess understanding of key concepts like Big O notation and differences between time and space complexity. This helps ensure that participants grasp the foundational elements before moving on to more complex topics.

2. Practical Coding Tests
If applicable, include practical coding tests where participants must write or identify the most efficient algorithms for given problems. This tests their ability to apply theoretical knowledge in practical scenarios.

3. Project-Based Assessment
Assign a small project where participants must analyse a system or software's performance and suggest improvements based on algorithmic complexity. This assesses their ability to apply their learning to real-world systems.

4. Group Presentations
Have participants work in groups to prepare presentations on how algorithmic complexity affects different sectors such as finance, healthcare, or technology. This evaluates their understanding and ability to communicate complex information.

5. Reflective Essays
Ask participants to write reflective essays on how the understanding of algorithmic complexity can impact their work or studies. This helps assess their ability to integrate and reflect on the knowledge gained.

Using these strategies, trainers can effectively deliver complex content on algorithmic complexity while ensuring participants are engaged, understand the material thoroughly, and can apply their knowledge practically.

5. References

Aksoy, T., & Gurol, B. (2021). Artificial intelligence in computer-aided auditing techniques and technologies (CAATTs) and an application proposal for auditors. In *Auditing Ecosystem and Strategic Accounting in the Digital Era: Global Approaches and New Opportunities* (pp. 361-384). Cham: Springer International Publishing.

Bentley, P. (1999). *Evolutionary design by computers*. Morgan Kaufmann.

Bohr, A., & Memarzadeh, K. (2020). The rise of artificial intelligence in healthcare applications. In *Artificial Intelligence in healthcare* (pp. 25-60). Academic Press.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). *Introduction to Algorithms*. MIT Press.

Davenport, T. H., & Mittal, N. (2023). *All-in on AI: How smart companies win big with artificial intelligence.* Harvard Business Press.

Yeung, K. (2018). Algorithmic regulation: A critical interrogation. *Regulation & governance, 12*(4), 505-523. https://doi.org/10.1111/rego.12158

Donovan, J., Caplan, R., Matthews, J., & Hanson, L. (2018). *Algorithmic accountability: A primer*. https://apo.org.au/node/142131

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., ... & Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management, 57*, 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

Felzmann, H., Villaronga, E. F., Lutz, C., & Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society, 6*(1), 2053951719860542. https://doi.org/10.1177/2053951719860542

Karp, R.M. (2010). Reducibility Among Combinatorial Problems. In: Jünger, M., et al. *50 Years of Integer Programming 1958-2008*. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-68279-0_8

Knuth, D. E. (1997). *The Art of Computer Programming: Fundamental Algorithms, volume 1.* Addison-Wesley Professional.

Mourtzis, D., Angelopoulos, J., & Panopoulos, N. (2022). A Literature Review of the Challenges and Opportunities of the Transition from Industry 4.0 to Society 5.0. Energies, 15(17), 6276. https://doi.org/10.3390/en15176276

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown.

Parker, G. G., Van Alstyne, M. W., & Choudary, S. P. (2016). *Platform revolution: How networked markets are transforming the economy and how to make them work for you.* WW Norton & Company.

Patel, K. (2024). Ethical reflections on data-centric AI: balancing benefits and risks. *International Journal of Artificial Intelligence Research and Development, 2*(1), 1-17. https://orcid.org/0009-0005-9197-2765

Qin, X. (2012). Making use of the big data: next generation of algorithm trading. In *Artificial Intelligence and Computational Intelligence: 4th International Conference*, AICI 2012, Chengdu, China, October 26-28, 2012. Proceedings 4 (pp. 34-41). Springer Berlin Heidelberg.

Reisdorf, B. C., & Blank, G. (2021). Algorithmic literacy and platform trust. In *Handbook of digital inequality* (pp. 341-357). Edward Elgar Publishing.

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., & Floridi, L. (2021). The Ethics of Algorithms: Key Problems and Solutions. In: Floridi, L. (eds) Ethics, Governance, and Policies in Artificial Intelligence. Philosophical Studies Series, vol 144. Springer, Cham. https://doi.org/10.1007/978-3-030-81907-1_8

CU4 | Data fairness and bias

Index

1. Introduction

Artificial Intelligence (AI) systems are not just abstract concepts but are increasingly integral to our daily lives. They influence decisions in sectors from healthcare and education to finance and security. However, as AI's influence grows, so does the potential for these systems to perpetuate existing societal biases or introduce new ones. Understanding the concepts of fairness and bias in AI is not just an academic exercise but crucial to developing just and equitable systems that impact our lives.

This CU on AI fairness and bias is about understanding the problem and empowering you to be part of the solution. It seeks to illuminate how biases can inadvertently arise in AI algorithms, the potential impacts of these biases, and strategies for mitigating them. The course also explores the ethical, societal, and technical challenges of creating fair AI systems, emphasizing your role in ethical AI development practices. Upon completing this CU, you will be able to:

Identify and Understand Types of Bias:

Recognize different types of biases that can occur in AI systems, including data bias, algorithmic bias, and outcome bias, and understand the mechanisms by which these biases manifest.

Measure and Quantify Bias:

Understand tools and methodologies to assess and measure the impact of biases within AI systems, using a range of fairness metrics and analytical techniques.

Develop Mitigation Strategies

Acquire practical knowledge on how to implement strategies to address and reduce bias in AI systems, including diverse data collection practices, algorithmic adjustments, and deployment monitoring.

Evaluate AI System Fairness

Should be able to critically evaluate AI systems for fairness across multiple dimensions, ensuring these systems do not perpetuate or exacerbate social inequalities.

Prepare for Future Challenges:

Able to anticipate and respond to emerging challenges in AI fairness as technology evolves, ensuring they remain adaptable and forward-thinking in their approaches.

TIPS AND RECOMMENDATIONS FOR TEACHERS

Suggestion: begin with a brief activity in which participants share their perceptions or personal experiences with AI in their daily lives. This can help participants connect the abstract concept of AI with its practical implications.

2. Fairness & bias - the candy analogy

IMAGE SOURCE | Generated by DALL-E

Imagine yourself in a classroom, eyeing a bowl of your favourite candy on the teacher's desk. Every student gets to pick a piece, but there's a twist: students wearing red shirts get two pieces while everyone else gets one. That doesn't seem fair. This is called 'bias'—when someone (or something, like an AI) gives unfair advantages to specific people for no good reason.

Now, imagine if, instead, the teacher asks everyone to solve a puzzle, and whoever gets it right, regardless of their shirt colour, gets an extra piece of candy. This feels fair because everyone has the same chance—it's about the puzzle, not the shirt. This is 'fairness,' where everyone has an equal opportunity, and no one is favoured for arbitrary reasons.

Like in this candy scenario, AI systems in the real world should give everyone a fair chance, whether recommending movies, approving loans, or playing music. When AI is fair, everyone solves the puzzle and gets the candy.

When it's biased, it's like some people are getting more candy just because of their shirt colour. And we want to aim for fairness!

Suggestion: employ this analogy interactively by having participants simulate the scenario with actual candies or a virtual activity. This hands-on approach could help solidify the concepts of fairness and bias.

3. Understanding the landscape: AI fairness and bias

Data Fairness in artificial intelligence refers to the principle that AI systems should treat all individuals equitably, without unjust or prejudiced outcomes based on inherent or acquired characteristics. Fairness in AI ensures that the decisions made by these systems do not favour or disadvantage any particular group of users over others. This is particularly important in applications affecting people's lives, such as hiring, loan approvals, law enforcement, and healthcare diagnostics.

Bias in AI, meanwhile, refers to systematic errors in the data or algorithms that lead to unfair outcomes for specific groups. Biases can be explicit, such as those intentionally built into a system, or more often, implicit, arising from unrepresentative or incomplete training data or flawed algorithmic design that inadvertently advantages or disadvantages specific populations.

IMAGE SOURCE | Generated by DALL-E

Addressing fairness and bias in AI development cannot be overstated. AI systems that operate with bias can perpetuate and amplify existing social inequalities, leading to hard-to-break cycles of disadvantage. Moreover, biased systems erode trust in technology, potentially stalling adoption and innovation. Ensuring fairness

in AI is a technical necessity and a moral imperative to maintain societal trust and prevent harm.

TIPS AND RECOMMENDATIONS FOR TEACHERS

Suggestion: use case studies or news articles highlighting recent issues of bias in AI systems, followed by group discussions. This can contextualize the importance of fairness in AI.

Gender Bias in Job Recruitment AI

In 2018, it was revealed that an AI system used by Amazon for recruiting was biased against women. The AI had taught itself that male candidates were preferable by observing patterns in resume submissions over a 10-year period, predominantly from men, due to the tech industry's gender imbalance. The system downgraded resumes that included the word "women's," as in "women's chess club captain" or "women's college."

4. Bias in AI systems

Bias in AI refers to systematic errors that result in unfair outcomes for certain groups, typically favouring one demographic over others. These biases can arise from various sources, such as the data used to train AI systems, the algorithms that process this data, or the interpretative biases of those who design and deploy these systems.

4.1. Occurrence of Bias in AI Systems

- **Data Bias**: This bias occurs when the datasets used to train AI algorithms do not represent the broader population. This can happen due to historical inequalities or simply because the data collection methods capture one demographic more thoroughly than others.
- **Algorithm Bias**: Sometimes, the algorithms may be designed in ways that inherently favour specific outcomes. This can be due to how algorithms interpret data, or the objectives set during AI development.
- **Feedback Loops**: Bias can also be perpetuated and amplified by feedback loops where initial biased decisions lead to further data collection that reinforces these biases, creating a cycle of discrimination.

4.2. Consequences of Bias in AI

- **Social Impact**: AI bias can reinforce existing social inequalities and stereotypes, leading to a society where digital technologies systematically disadvantage certain groups. This erosion of trust in AI technologies can have broader implications for the adoption and effectiveness of AI across various sectors.

- **Individual Impact**: On a personal level, bias in AI can lead to unfair treatment in critical areas such as hiring, law enforcement, healthcare, and financial services. Individuals may face unjust outcomes due to flawed decision-making processes influenced by biased AI systems.

4.3. Addressing AI Bias

Adopting comprehensive strategies throughout the AI development lifecycle to mitigate AI bias is essential. These include:

- Diversifying data sources to ensure broader representation in training datasets.
- Developing and implementing algorithmic audits to identify and correct biases.
- Establishing continuous monitoring systems to detect and address emergent biases in deployed AI systems.

4.4. Overview of Types of Bias

Understanding the different types of biases that can manifest in AI systems is essential for developers, policymakers, and users who aim to create and interact with fair and equitable technologies. This section will explore the various forms of bias, each accompanied by real-world examples illustrating their impact.

1. Selection Bias: Selection bias occurs when the data used to train an AI system do not represent the target population or scenario, leading to skewed results. For example, an AI model trained to recognize facial features using a dataset primarily composed of younger individuals may fail to accurately identify features in older adults, as the model's training did not include a representative sample of all age groups.

2. Confirmation Bias: Confirmation bias in AI manifests when data or its interpretation by model designers inadvertently supports pre-existing beliefs, ignoring contradictory evidence. An AI used in hiring trained on resumes predominantly from one gender may continue to favor candidates of that gender, reinforcing the existing imbalance in job selections.

3. Sampling Bias: Sampling bias is a specific type of selection bias where the sample data collected to train AI does not accurately represent the broader population. If a voice recognition system is trained mainly on data from speakers with American accents, it might struggle to understand accents from other

regions, such as British or Australian, due to its training on a non-representative sample.

4. Algorithmic Bias: Algorithmic bias occurs when the algorithms that underlie AI systems produce biased outcomes, often due to inherent flaws in their design or the objectives set during their creation. A credit-scoring AI that disproportionately denies loans to applicants from certain neighbourhoods due to historical data reflecting past lending biases exemplifies algorithmic bias.

5. Reporting Bias: Reporting bias arises when the data fed into an AI system are skewed by the tendency to report only certain kinds of outcomes. For instance, an AI monitoring system for drug side effects that primarily receives reports of severe cases may erroneously suggest that most patients experience extreme reactions, ignoring milder but more common side effects.

6. Measurement Bias: Measurement bias involves errors in data collection that systematically skew the data. An environmental monitoring AI that relies on less effective sensors in colder temperatures will have a biased view of pollution levels, potentially under-reporting emissions during winter months.

5. Methods for identifying and measuring biases in AI

Understanding and mitigating biases in artificial intelligence are critical to building fair and effective systems. This section details methodologies for identifying and measuring these biases, providing a clear pathway for AI developers and data scientists to ensure their AI systems operate without unfair biases.

- **Data Auditing** is a comprehensive review process where datasets used to train AI models are scrutinized for potential biases or anomalies. To perform a data audit, one should:
    - **Assess Data Representativeness**: Evaluate whether the dataset accurately reflects the diversity of the population the model will serve. This includes checking demographic representativeness and scenario coverage.
    - **Identify Missing Data**: Look for patterns in missing data — gaps can introduce biases if specific subgroups are underrepresented.
    - **Analyse Label Consistency**: Ensure the data labelling process is consistent and unbiased. Labelling discrepancies can lead to significant biases in model outcomes.
    - **Statistical Analysis**: Use statistical tools to detect anomalies or skewed data distributions that might indicate biased data.
- **Disparity Analysis** involves comparing model performances across different demographic groups or other relevant subgroups to identify discrepancies that might indicate bias. The steps to conduct a disparity analysis include:

- **Define Relevant Subgroups:** Based on attributes like age, gender, ethnicity, or other characteristics pertinent to the model's application.
- **Measure Performance Metrics:** Calculate performance metrics such as accuracy, precision, recall, and F1-score for each subgroup.
- **Compare Outcomes:** Analyse the differences in performance metrics between groups. Significant discrepancies may indicate the presence of bias.
- **Root Cause Analysis:** If disparities are found, investigate the underlying causes, which may relate to data collection, model architecture, or feature selection.

- **Sensitivity Analysis** tests an AI model's robustness by altering input data slightly and observing how the outputs change. This is particularly useful for identifying biases:
  - **Vary Input Data**: Introduce small changes to input data that reflect plausible variations in real-world scenarios.
  - **Monitor Output Fluctuations**: Observe whether these changes lead to disproportionate changes in the model's outputs.
  - **Assess Impact on Specific Groups**: Focus on how these changes affect model outputs for different demographic groups.
  - **Implement Adjustments**: Use findings to refine data preprocessing, feature engineering, or the model to reduce unwanted sensitivities.

6. Strategies for Addressing and Mitigating Biases

Mitigating biases in AI systems involves a multi-faceted approach during the design, development, and deployment phases. The following strategies are essential to **reduce the risk of biased outcomes**.

- **Diverse Representation** Ensuring diversity in data and team composition is crucial for developing unbiased AI systems.
  - **Data Diversity**: Collect data from a wide array of sources to cover a broad spectrum of the population. This includes different demographics, socio-economic backgrounds, and other relevant characteristics.
  - **Team Diversity**: Diverse teams bring various perspectives that can identify and mitigate potential biases that might not be evident to a more homogeneous group.
- **Bias-Aware Algorithms** Developing algorithms that are inherently aware of and can adjust for biases is a proactive approach to mitigating biases.

- **Incorporate Fairness Metrics**: Integrate fairness considerations into the algorithm's objective function, such as minimizing disparity in error rates across groups.
- **Adversarial Training**: Implement adversarial training techniques where models are simultaneously trained to predict correctly and to minimize bias.

- **Regular Monitoring and Evaluation** Continuous monitoring and regular evaluation of AI systems are vital to ensure they remain unbiased over time and adapt to new data or contexts.
  - **Establish Monitoring Frameworks**: Set up systems that continuously track the performance of AI models, primarily focusing on fairness metrics.
  - **Periodic Reviews**: Regularly review and recalibrate the models based on ongoing monitoring results, adapting to changes in data patterns or societal norms.
  - **Stakeholder Feedback**: Engage with stakeholders, including end-users, to gather feedback on AI performance and perceived fairness, which can provide insights not captured in quantitative evaluations.

7. Understanding Fairness Metrics

Fairness metrics are crucial for evaluating how well AI systems treat individuals or groups from different backgrounds. They provide a framework to measure and ensure that AI systems do not perpetuate or exacerbate existing societal biases. Below are some of the **critical fairness metrics** that should be considered:

- **Statistical Parity (Demographic Parity)**: This metric assesses whether the proportion of positive outcomes is the same across different demographic groups (e.g., gender, race). It is crucial for applications where the goal is to ensure equality of outcome regardless of input features related to group identity.
- **Equalized Odds**: This metric requires that the true positive rate (sensitivity) and false positive rate (specificity) are the same across groups. It is used primarily in settings where the consequence of an error affects all groups equally, such as in criminal justice or hiring.
- **Conditional Demographic Disparity**: This metric evaluates differences in predictive outcomes conditioned on legitimate attributes. It measures disparities not justified by relevant attributes and is suitable for scenarios where some differences between groups are expected due to those attributes.

- **Fairness Through Awareness**: This approach involves incorporating awareness of sensitive attributes (like race or gender) directly into the decision-making process to ensure equitable decisions. This metric advocates for algorithms that compensate for historical injustices or societal biases that could affect fairness.
- **Individual Fairness:** This metric asserts that similar individuals should receive similar predictions regardless of their group membership. This concept is especially relevant in personalized services like job recommendations, where consistency in treatment across similar user profiles is essential.
- **Counterfactual Fairness**: Under this metric, a decision is considered fair if it would have been made in a counterfactual world where the individual belonged to a different demographic group but was otherwise identical. This metric is valuable in assessing individual-level fairness and helps understand the impact of specific attributes on decisions.

Selecting appropriate fairness metrics depends on the specific context and goals of the AI application. Here are some guidelines to consider:

1. **Application-Specific Requirements:** The choice of fairness metric should align with the goals and legal requirements of the application domain. For example, Equalized Odds may be crucial for criminal justice applications, while Statistical Parity might be more appropriate for marketing or advertising.
2. **Regulatory and Ethical Considerations:** Some industries may have regulatory requirements that dictate specific fairness metrics. Ethical considerations, such as the need to redress historical injustices, can also guide the choice of fairness metrics.
3. **Impact Assessment:** It's essential to consider the potential social impact of the AI system and choose metrics that minimize harmful biases. For instance, Counterfactual Fairness can be crucial in domains like healthcare, where individualized treatment is necessary.
4. **Technical Feasibility:** Depending on the AI system's complexity and data availability, some metrics may be more challenging to implement than others. Technical limitations should be considered to ensure that the chosen fairness measures can be accurately and consistently applied.

8. Fairness in Data Collection and Preprocessing

Ensuring that data collection practices are diverse and unbiased is fundamental to building fair AI systems.

- **Representative Sampling**: Make sure that the data collection process includes samples from all population segments that the AI system will serve. This may involve stratified sampling to include underrepresented groups.
- **Ongoing Data Collection**: Regularly update and expand data collections to reflect new and emerging trends within the population, as static data sets can quickly become outdated.

8.1. Preprocessing Techniques

Ensuring fairness in data collection and preprocessing is essential to prevent biases from seeping into AI systems. Below are several practices that can help achieve a more balanced and equitable AI system through careful attention to the initial stages of AI development.

- **Handling Missing Data**: Analyse patterns of missing data as these can introduce bias. Techniques like imputation should be carefully applied to avoid introducing additional biases.
- **Feature Selection**: Critically assess which features are included in the model training process. Features correlated with sensitive attributes should be cautiously handled to avoid proxy discrimination.
- **Normalization and Standardization**: Apply normalization or standardization to ensure that the model does not inherently weigh certain features more heavily simply due to their numerical scale.
- **Diverse Data Sources:** Ensuring the data collection encompasses a broad spectrum of population demographics is critical to prevent biases in AI models. The goal is to gather data that reflects the variety and diversity of the real world where the AI system will operate. This involves engaging with various demographic groups and utilizing multiple collection points. For example, when developing a healthcare AI, data should be collected from hospitals across different regions, including rural and urban areas, to capture diverse health profiles and environmental impacts on diseases.
- **Inclusive Sampling**: Inclusive sampling aims to ensure that all population segments are represented in the training dataset. The objective is to prevent the exclusion of minority or underrepresented groups whose absence could lead to biased AI predictions. Techniques like stratified sampling are used where the population is divided into subgroups (strata) that reflect key characteristics (e.g., ethnicity, age, gender), and samples are randomly selected from each stratum to ensure inclusivity in the dataset.
- **Data Cleaning**: Data cleaning involves removing inaccuracies and inconsistencies that can skew AI model outcomes. The objective is to ensure the data quality is high and free of errors that could bias the model's decisions. This process includes correcting or removing outliers,

filling missing values using appropriate imputation strategies that do not bias the data, and ensuring all data entries are consistent and formatted correctly.

- **Regular Monitoring**: Monitoring data and model performance is crucial to detect and address any biases that may emerge over time. The goal is to maintain ongoing vigilance to ensure that the AI system continues to operate fairly as new data is collected and as conditions change. Setting up automated monitoring systems that continuously assess the AI outputs for fairness metrics and alert developers if biases are detected. This system should regularly evaluate the data feeding into the AI to check for data quality or representation changes.

- **Fair Feature Engineering:** Fair feature engineering involves selecting and transforming features to minimize biases while preserving the data's utility. The objective is to ensure that model inputs do not inherently disadvantage any group. Techniques include analysing the correlation of features with sensitive attributes and either modifying or removing features that could lead to biased outcomes. Dimensionality reduction techniques can also identify and remove redundant or irrelevant features that may carry biases.

- **Transparency and Documentation:** Maintaining transparency through detailed documentation of the data collection, preprocessing, and feature selection processes. The objective is to provide clear records that can be audited for compliance with fairness standards. Keeping detailed logs of data handling decisions, methodologies used for data cleaning, and the rationale behind feature selection. These documents should be readily available to stakeholders and regulatory bodies to review and assess the fairness of the AI system.

- **Bias Detection**: Bias detection involves identifying potential biases in the dataset before it is used to train AI models. The objective is to correct any biases affecting the model's fairness pre-emptively. Employing statistical analysis and visualization tools to examine data distributions across different demographic groups. This can involve specialized software, or algorithms designed to highlight areas where data may not be representative or where outcomes may disproportionately affect certain groups.

The fairness of an AI system starts with how data is collected, processed, and prepared before it even reaches the algorithmic training stage. By implementing rigorous practices in data handling and preprocessing, developers can significantly reduce the risk of biases in AI systems. These practices not only enhance the fairness and reliability of AI applications but also build trust among users and stakeholders in diverse and regulated environments.

9. Emerging Trends in AI Fairness

As AI technologies evolve, they encounter complex datasets and dynamic regulatory environments. New technologies like explainable AI (XAI) and federated learning offer opportunities to enhance AI fairness by providing greater transparency into AI decision-making processes and allowing for decentralized data processing that can help protect individual privacy. Future challenges include:

- **Complex Data Interactions**: With the advent of big data, understanding the interactions and correlations within large and complex datasets becomes more challenging but essential for ensuring fairness.
- **Adaptive Regulations**: As public awareness of AI biases increases, regulatory bodies will likely introduce more stringent fairness guidelines, requiring AI systems to be adaptable and transparent about their fairness measures.
- **Explainability and Interpretability**: Explainability and interpretability in AI involve making the operations and decisions of AI systems transparent and understandable to humans. The objective is to ensure stakeholders comprehend how AI decisions are made, which is critical for trust and accountability. This trend involves developing methods and tools that allow for the visualization and explanation of AI decision processes. Techniques like feature importance scores, decision trees, and model-agnostic methods provide insights into complex models' workings, such as deep neural networks.
- **Adversarial Attacks and Defenses**: Adversarial attacks involve manipulating AI inputs to deceive AI systems into making incorrect decisions. The defense against such attacks aims to make AI systems more robust and secure and protect them from malicious inputs that could exploit biases or weaknesses. Implementing defenses against adversarial attacks includes using techniques like adversarial training, where the model is exposed to adversarial examples during training to learn to resist them. Other approaches involve regular security assessments and updating models to recognize and counteract new attack strategies.
- **Fairness Across Multiple Dimensions:** Fairness across multiple dimensions refers to addressing biases that intersect with various demographic and socio-economic factors, including, but not limited to, race, gender, age, and disability. The objective is to ensure comprehensive fairness that accounts for complex human identities and their interactions. This involves developing multidimensional fairness metrics and using intersectional data analysis techniques to assess and mitigate biases at the intersection of multiple attributes.

- **Fairness in Emerging Technologies**: As AI is integrated into emerging technologies like autonomous vehicles, smart healthcare, and IoT devices, ensuring fairness in these applications becomes crucial. The objective is to implement fairness from the ground up in these new technologies to prevent biases that could have widespread societal impacts. This involves conducting impact assessments and bias audits for new technologies and designing fairness guidelines specific to each technology's context and use case.

- **Ethical Consideration and Governance**: Ethical consideration and governance in AI involve establishing frameworks and guidelines that ensure AI systems are developed and deployed in a manner that upholds ethical standards and societal values. The objective is to provide a regulatory and ethical oversight mechanism to guide AI development and usage. Developing comprehensive AI ethics guidelines, forming ethics boards, and engaging with policymakers to enact regulations that enforce fairness, transparency, and accountability in AI.

- **Human-Centric Design**: Human-centric design creates AI systems prioritizing human welfare, values, and ethical considerations. The objective is to ensure that AI technologies enhance human capabilities and quality of life without replacing or diminishing human input and oversight. This involves engaging with diverse user groups during the design process, incorporating feedback loops that allow users to inform ongoing AI development, and ensuring that AI supports rather than supplants human decision-making.

- **Algorithmic Bias Mitigation:** Algorithmic bias mitigation involves identifying and correcting biases in AI algorithms to prevent unfair outcomes. The objective is to refine AI models continuously to ensure they are as objective and unbiased as possible. This includes using bias detection and correction techniques during the model training process, such as re-weighting training data, adjusting model parameters to balance accuracy and fairness, and employing external bias mitigation algorithms.

These trends and issues highlight the dynamic nature of AI development and the ongoing need for innovation and vigilance in ensuring fairness as AI systems become more integrated into every aspect of human life.

10. Conclusion

Exploring AI fairness and bias is crucial as we integrate artificial intelligence into more facets of daily life and critical infrastructure. From the definitions and impacts of various biases in AI systems to the strategies for their identification, measurement, and mitigation, it is evident that fairness in AI is a multifaceted challenge requiring comprehensive and ongoing attention.

IMAGE SOURCE | Generated by DALL-E

- **Understanding Bias Types and Impacts:** The first step towards mitigation is to recognize different types of biases, such as selection, algorithmic, and confirmation biases, and understand their potential to perpetuate inequality. Real-world examples like facial recognition inaccuracies and biased hiring algorithms have illustrated the detrimental effects of unchecked AI biases.

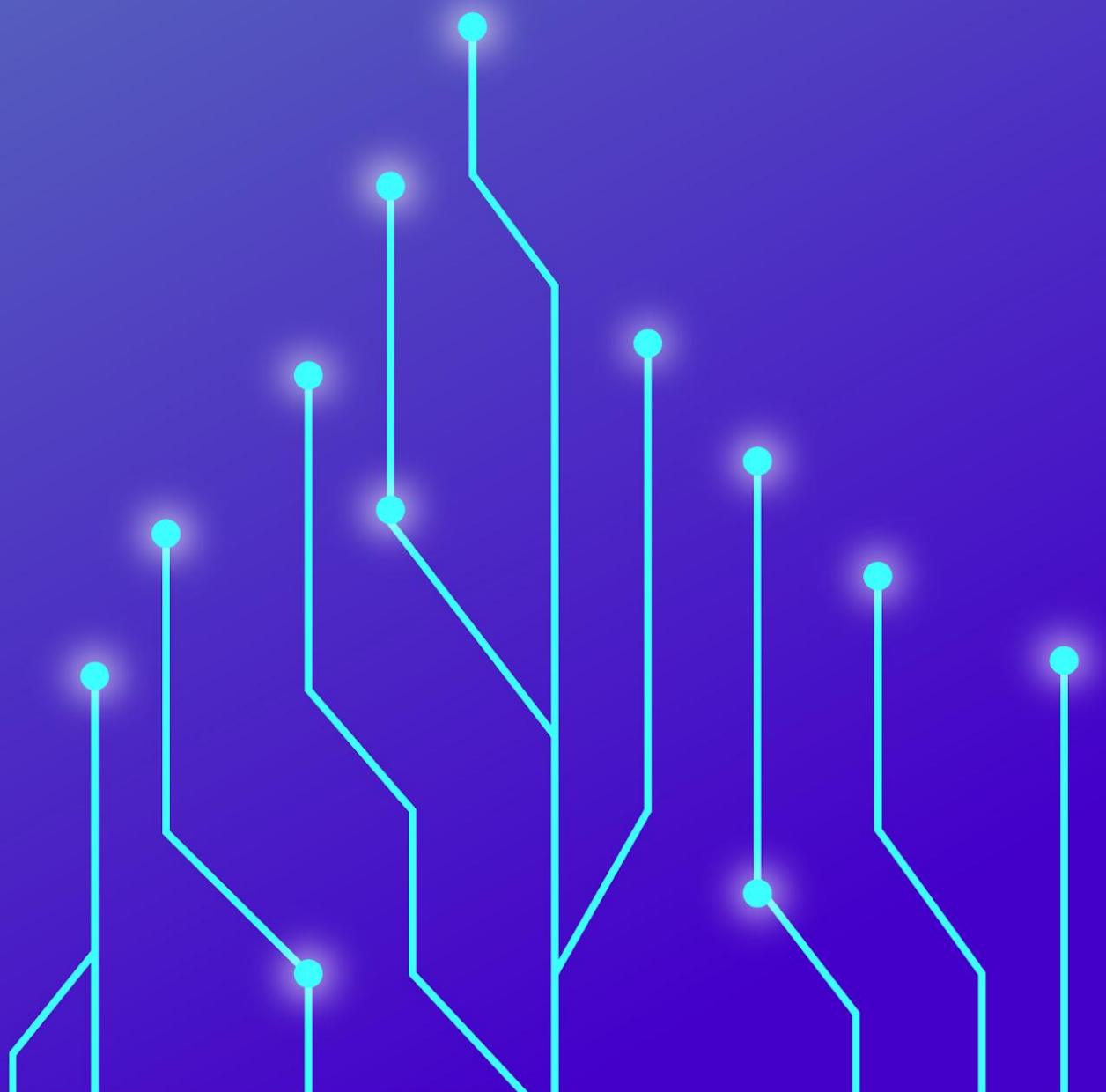
- **Strategies for Fairness:** Effective methods to ensure fairness include diverse data collection, inclusive sampling, and regular monitoring of AI systems. These strategies help create AI models that are representative and fair and reduce systematic errors that disadvantage certain groups.

- **Emerging Trends and Challenges:** AI fairness is evolving, with new challenges such as ensuring fairness in emerging technologies and developing defenses against adversarial attacks. There is also a growing emphasis on explainability and human-centric design, which are key to building trustworthy AI systems.

- **Ethical Considerations and Governance:** The importance of ethical governance in AI development cannot be overstated. Establishing robust ethical guidelines and governance frameworks will be critical in guiding the responsible deployment of AI technologies.

- **Continuous Learning and Adaptation:** As AI technologies and societal norms evolve, so must our approaches to fairness. Developers, policymakers, and stakeholders need continuous education, vigilance, and adaptation to ensure that AI systems remain fair and beneficial to all.

CU5 | Case studies and projects



INDEX

1. Introduction

In this course unit, the focus is on examining biases in a practical context. The project involves a hypothetical scenario in which an international company develops an AI solution for recruiting. This project will be addressed during the final interactive study session with students. Please facilitate a lively discussion to support their learning, encouraging students to express and apply what they have learned in previous sessions.

Algorithmic bias is the production of unfair or discriminatory outcomes by algorithmic systems. This occurs when an algorithm systematically favours or disadvantages certain individuals or groups based on specific characteristics, such as race, gender, or socioeconomic status. Machine learning algorithms operate on various datasets, such as training data, from which they learn models applicable to other individuals or groups for making predictions. However, these algorithms might treat similar individuals or items differently, leading to the replication or magnification of human biases, particularly those affecting protected groups (Friedman & Nissenbaum, 1996; Lange & Duarte, 2018; Lee et al., 2019). Understanding the causes and implications of algorithmic bias is crucial for developing ethical and trustworthy AI solutions.

Implications of algorithmic bias

Algorithmic bias can appear in various forms, affecting different groups in distinct ways. The examples collated by Lee, Resnick, and Barton (2019) illustrate how bias can originate from multiple sources and impact groups in unintended or deliberate ways.

- Word associations can reflect bias. Princeton University researchers found that European names were perceived more positively than African American ones, and words like “woman” and “girl” were linked more to arts than to science and math (Lee et al., 2019). As demonstrated in the initial example of this course, associations like 'nurse' with women and 'engineer' with men are often shaped by biased data embedded in search algorithms, reflecting societal stereotypes. This perpetuates bias in AI and machine learning due to the reliance on existing online content, posing a significant ongoing challenge in the field of AI and machine learning.
- In online recruitment tools, Amazon discontinued the use of a recruiting algorithm due to gender bias. The AI software penalized resumes containing the word “women’s” and downgraded those from women's colleges.
- Online ads also exhibit bias. Harvard researcher Latanya Sweeney discovered that searches for African American names were more likely to return ads for arrest record services compared to white names.

- Facial recognition technology can be biased too. MIT researcher Joy Buolamwini found that commercial systems often fail to recognize darker skin tones, as they are trained predominantly on lighter-skinned and male faces.
- Even algorithms in criminal justice can be biased. The COMPAS algorithm used by judges to predict bail decisions was found to be biased against African Americans, as reported by ProPublica.

2. Different types of bias

According to Friedman and Nissenbaum (1996), algorithmic biases can be categorized into three main types.

- First, **pre-existing bias** refers to biases that already exist within a system, whether intentionally or unintentionally, before its creation. These biases can enter a system through deliberate efforts or unconscious actions, sometimes even with good intentions.
- Second, **technical bias** arises from technical constraints or considerations during the design process. Various aspects of design can contribute to technical bias.
- Third, **emergent bias** occurs in the context of real-world usage with actual users. This type of bias typically emerges after the design process is completed, often due to changes in societal knowledge, population or cultural values. (Friedman & Nissenbaum, 1996)

Examples of Bias in Machine Learning Data Sets

Machine learning models heavily rely on training data, but the human involvement in selecting and preparing this data can introduce various types of bias. One prevalent example is **reporting bias**, where the frequency of events in the dataset doesn't mirror real-world occurrences. This bias can lead to models struggling to handle ordinary situations because of an overemphasis on documenting unusual circumstances. (Google for Developers, 2022)

Another type is **group attribution bias**, which manifests in favouring one's own group or stereotyping others based on group characteristics. Implicit bias, stemming from personal experiences and mental models, further complicates matters, sometimes leading to **confirmation bias** or **experimenter's bias** during model training. (Google for Developers, 2022)

Automation bias, where automated systems are favoured over human judgment regardless of their accuracy, and **selection bias**, arising from non-representative data collection methods, also pose significant challenges in mitigating bias in machine learning models. (Google for Developers, 2022)

Real-world case examples

After understanding the theoretical framework of algorithmic bias and its implications, it is now appropriate to explore real-world case examples that illustrate the impacts of algorithmic bias. The first case delves into TikTok's harmful algorithm, while the other explores ethical concerns surrounding facial recognition technology.

TikTok's harmful algorithm

In a revealing investigation by Finland's national media company Yle, the harmful effects of TikTok's algorithm came to light. Yle conducted a study focusing on the content delivered to a fictional 13-year-old girl named Ella, who suffered from depression and body image issues. (YLE, 2023)

In their study, Yle's data scientists navigated TikTok using Ella's profile for five hours. Initially, Ella encountered cheerful videos ranging from food and pets to jokes. However, within minutes, the algorithm began to serve content related to mental health, including antidepressants. As the browsing continued, the tone of the videos grew darker, with TikTok displaying content on suicide, self-harm, and eating disorders. By the end of the study, a staggering 95 % of the content presented to Ella was deemed harmful, including glorification of eating disorders and advice on restricting food intake. (YLE, 2023)

Social psychologist Suvi Uski highlighted the concerning aspect of the algorithm's behaviour, emphasizing its tendency to prioritize user engagement over the potential harm caused by the content shown. This revelation underscores the significant ethical implications associated with TikTok's algorithmic recommendations. (YLE, 2023)

What began as innocent browsing quickly escalated into a concerning spiral of harmful content, reflecting Ella's vulnerable state. The algorithm's rapid identification of Ella's interest in depressive and unhealthy behaviours is alarming in itself, but even more so is its role in amplifying these harmful thoughts rather than offering a counterbalance. Instead of redirecting Ella towards positive and supportive resources, the algorithm reinforced her negative tendencies, exacerbating her mental health struggles. This raises critical questions about the ethical responsibility of algorithm designers and platform operators in safeguarding user well-being. The case of Ella underscores the urgent need for greater transparency, accountability, and ethical considerations in the development and deployment of algorithmic systems, particularly those that wield significant influence over user behaviour and mental health.

Unethical facial recognition data collection

Facial recognition technology has become increasingly prevalent in various sectors, from healthcare to law enforcement (Javaid, 2024). However, its widespread adoption has raised significant ethical concerns, particularly regarding data collection and its impact on privacy and civil liberties. A notable example of unethical facial recognition data collection surfaced in a Washington Post report on July 7, 2019. The report revealed that federal agencies like the FBI and ICE had accessed driver's license databases for facial recognition purposes, scanning through millions of Americans' photos without their consent or knowledge (Harwell, 2019a).

This alarming revelation underscores the broader issue of facial recognition technology's potential for bias and discrimination. A subsequent federal study, also reported by The Washington Post on December 19, confirmed the prevalence of racial bias within facial recognition systems. The study found that Asian and African American individuals were up to 100 times more likely to be misidentified than white men, depending on the algorithm and type of search used (Harwell, 2019b).

3. Tools to Assess the Ethics of an AI Solution

In the development of AI systems, particularly in areas sensitive to biases like recruitment, it's essential to employ various tools designed to identify and assess potential biases. These **assessment tools** are critical in ensuring that AI applications operate fairly and transparently. By utilizing such resources, developers can better understand how biases might influence system outcomes and take proactive steps to mitigate these effects. This approach not only improves the reliability and fairness of AI systems but also helps in building trust with users by demonstrating a commitment to ethical AI practices.

Design tools are crucial for ethical considerations in AI development because they facilitate a holistic understanding of various stakeholders' perspectives. These tools enable developers to empathetically consider the diverse impacts of AI technologies on different user groups. By mapping out the interactions, behaviours, and needs of these groups, these tools help to identify and mitigate potential biases inherent in AI systems. The broad engagement not only ensures that AI solutions are more inclusive but also aligns product development with ethical standards by proactively addressing issues that could lead to unfair or discriminatory outcomes. In this project case, different tools are presented throughout the design process under a relevant stage.

While there is a plethora of assessment tools available for this purpose, this course focuses on presenting only three of them. The design tools used in this project are detailed in the project description found later in this handbook Chapter E. Identifying potential biases.

The UNESCO Ethical Impact Assessment tool, developed in 2023, offers a structured approach to evaluating the ethical implications of AI systems. It comprises a series of diverse questions, encompassing both open-ended and multiple-choice formats, designed to guide users through a comprehensive assessment process. Emphasizing ethical considerations from the outset, the tool begins with scoping inquiries before delving into UNESCO principles. By addressing key aspects such as stakeholder impacts, team responsibilities, safety, security, and concerns regarding sustainability, privacy, transparency, explainability, and accountability, it facilitates a thorough examination of the ethical dimensions inherent in AI deployment. (UNESCO, 2023)

The Fairness, Non-Discrimination, and Diversity section of the UNESCO Ethical Impact Assessment tool focuses on identifying algorithmic bias, especially in the data used by AI systems. It thoroughly examines the roles and responsibilities involved to ensure all inherent biases are addressed. (UNESCO, 2023)

The Assessment List for Trustworthy AI (ALTAI) by the High-level expert group on artificial intelligence (AI HLEG), established by the European Commission, provides a structured framework for evaluating the trustworthiness of AI systems. This tool outlines seven specific requirements for trustworthy AI, with detailed explanations available on the accompanying website. Additionally, it offers guidance on completing the assessment and identifies key stakeholders who should be involved in the process, ensuring a comprehensive evaluation of AI systems' trustworthiness. (ALTAI, 2020)

The ALTAI tool's section Diversity, Non-discrimination and Fairness focuses on avoiding unfair bias. It assesses algorithmic bias through targeted questions, evaluating input data and algorithm design for biases, while promoting inclusivity, accessibility, and stakeholder participation to ensure fairness and transparency in AI systems. (ALTAI, 2020)

Mario Sosa Hidalgo's Design of an Ethical Toolkit for the Development of AI Applications offers an innovative, concrete, and visual approach to integrate ethics into the development process of AI applications. The tool addresses bias and advises developers to implement strategies that minimize unfairness, ensuring the AI system is as unbiased as possible. (Hidalgo, 2019)

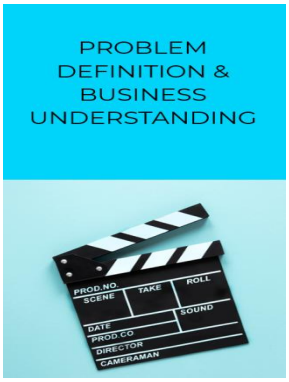
4. Project: AI solution in recruitment

This section focuses on addressing algorithmic bias within AI systems. It presents a project that involves a hypothetical scenario in which an international company develops an AI solution for recruiting.

To guide this exploration, the narrative is contained in text boxes outlined in blue, which serve as essential inputs for the project. Additionally, this section provides

comprehensive information to help understand the context and the development of the AI application for recruiting purposes.

The aim is to engage students in recognizing and mitigating biases that can arise during the creation and implementation of AI technologies.

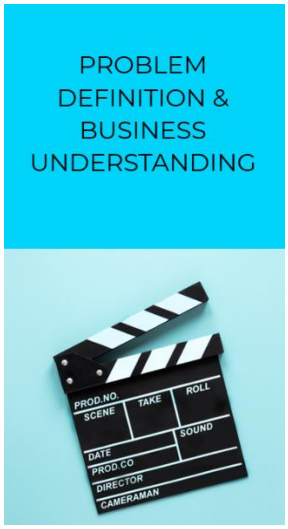


Setting the Scene

A global industrial company faces inefficiencies and uncertainties in its recruitment process, which puts a significant strain on the HR department by consuming time and resources.

Anticipating future growth, the company also recognizes the urgent need to hire a variety of specialists, including engineers, marketers, finance professionals, coders and project managers.

Artificial intelligence (AI) is increasingly being integrated into recruitment processes, offering a range of possibilities for improving hiring practices. While AI applications in recruitment bring significant advantages, they also present certain challenges. It is crucial for organizations to recognize and carefully consider both sides of the coin. By understanding the potential benefits as well as the drawbacks, companies can make informed decisions on how best to implement AI technologies. This balanced approach ensures that while harnessing the power of AI to streamline and enhance recruitment, the implications and risks are also adequately addressed.



Understanding the problem

The company begins background research to develop an AI solution, forming an internal team dedicated to identifying key challenges in the existing recruitment process. The team conducts interviews with a range of stakeholders, including the HR department, newly hired employees, and managers who have recently onboarded new team members. The team also gathers data on the current costs of recruitment and analyzes the success rate of these initiatives.

Understanding the perceptions of potential employees regarding the use of AI in recruitment is crucial for the company. It aims to avoid any damage to its reputation or brand. To achieve this, the company is committed to ensuring that the adoption of artificial intelligence not only gains widespread acceptance but also adheres to ethical and legal standards.

After discussions and investigations, the pros and cons of an AI recruitment solution are listed as follows:

Benefits of AI in recruitment

Artificial intelligence is revolutionizing the recruitment process by automating routine tasks, thereby allowing recruiters to concentrate on more strategic aspects of their work. This automation not only enhances the overall efficiency in handling applications but also significantly reduces the time spent on recruitment processes. (Albaroudi et al., 2024; www.recruiter.com, 2023)

Furthermore, AI's ability to analyse extensive datasets enables the identification of the best-fit applicants, thereby improving the quality of hires. This analytical capability potentially reduces turnover by ensuring a good match between the job and the candidate. The quick feedback and updates facilitated by AI not only enhance communication but also keep applicants well-informed throughout the selection process. This allows recruiters more time to engage with selected candidates, further improving the applicant experience. (Arivu Recruitment and Consulting, 2023; Sheard, 2022; www.recruiter.com, 2023)

The implementation of AI in recruitment also leads to cost reduction by minimizing the reliance on manual labour. (www.recruiter.com, 2023) Additionally, AI strives to diminish human biases, thus increasing objectivity, consistency, and fairness in the recruitment process. This technological advancement is particularly beneficial for groups facing barriers to economic participation, as it potentially improves their hiring prospects. (Bursell & Roumbanis, 2024; Köchling & Wehner, 2020; Sheard, 2022)

AI's capability for insightful analysis extends to social media screening, where it assesses candidates' profiles for cultural fit and job role suitability. It also screens for inappropriate content, ensuring that recruitment aligns with organizational

values. Moreover, AI tools aid in overcoming language barriers in recruitment, enabling access to a broader, more diverse talent pool globally. (Albaroudi et al., 2024)

Lastly, AI is expected to enhance workplace diversity by removing bias from the hiring process. (Sheard, 2022) This array of benefits underscores the transformative impact of AI on the recruitment landscape, promising a more efficient, fair, and inclusive hiring process.

Concerns of AI in recruitment

The introduction of AI in the recruitment process, while offering numerous efficiencies, brings with it the challenge of limited human touch and impersonal interaction. This reduction in personal engagement can make the recruitment experience feel mechanical, potentially alienating candidates and diminishing the quality of the recruitment experience. (www.recruiter.com, 2023)

Moreover, the risk of bias and discrimination poses a significant concern. AI, if not properly designed, trained, or if based on biased training data, can inadvertently introduce biases into the recruitment process. This can negatively impact diversity, inclusion, and fairness in hiring practices. (Albassam, 2023; www.recruiter.com, 2023)

Data protection and security concerns are also paramount, given the vast amounts of personal data collected and processed by AI systems. These concerns raise issues around privacy, data protection, and the risk of cyberattacks, underscoring the need for robust security measures and transparency in data usage. (Albassam, 2023; www.recruiter.com, 2023)

The implementation of AI in recruitment is not without its costs and resource intensiveness. For small businesses, in particular, the financial and resource investment required can be substantial, posing a barrier to entry and potentially limiting the adoption of AI in their recruitment processes. (www.recruiter.com, 2023)

There are also concerns regarding the lack of human judgment and potential accuracy issues. Sole reliance on AI, without human oversight, might result in overlooking qualified candidates due to incomplete data or the failure to recognize the transferability of skills. Furthermore, the accuracy of AI can be compromised by candidates who manipulate their data to appear more qualified. (Albassam, 2023; Arivu Recruitment and Consulting, 2023)

The evolving legal framework surrounding the use of artificial intelligence in recruitment requires companies to adapt and comply to ensure fairness. This evolving landscape highlights the need for constant vigilance and adaptability in the use of AI tools in the recruitment process. (Fernandes, 2021; www.recruiter.com, 2023)

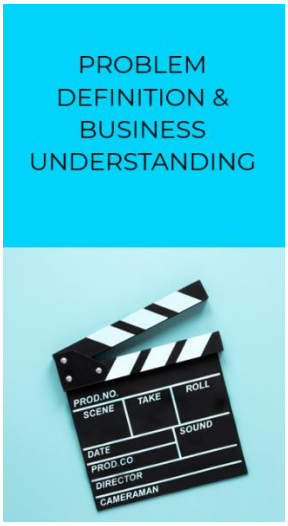
Limitations in AI's ability to assess cultural fit or teamwork skills may result in missed opportunities to hire candidates who could excel if given the chance. Albassam (2023) emphasizes the importance of considering these factors, which are often critical for job performance but difficult for AI systems to quantify accurately.

Market dominance and a lack of transparency in AI hiring tools, particularly those developed by private corporations in the US, complicate efforts towards bias mitigation and external scrutiny. The proprietary nature of these systems, as intellectual property, makes it challenging to evaluate and improve their fairness and effectiveness. (Sheard, 2022)

Lastly, the exacerbation of inequality through automated hiring platforms is a growing concern. These systems may reinforce existing social inequalities through biased algorithms and the significant role of networking and informal structures in job acquisition. This concern points to the need for careful consideration and action to remove biases from AI-powered hiring platforms. (Harvard John A. Paulson School of Engineering and Applied Sciences, 2023)

Each of these points underscores the complexity and multifaceted nature of incorporating AI into the recruitment process, highlighting the need for a balanced approach that considers efficiency, fairness, and inclusivity.

5. Identifying potential biases



Understanding the users of the system

One of the interviewed persons is Lead Recruiter Maria, who has several years of experience in the company. She has seen firsthand the accumulation of recruitment tasks and the growing pressure it causes on the HR department, which intensifies when hiring global talent.

Maria is open-minded and interested in the possibilities of AI to help with recruitment tasks, even though she is not deeply familiar with artificial intelligence technology.

Tool: Persona

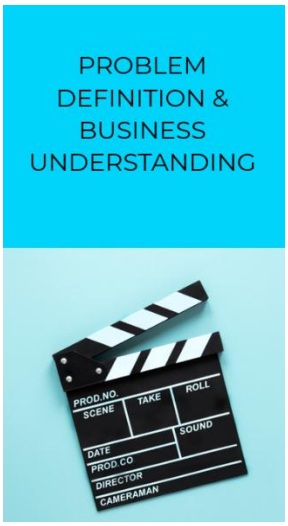
Personas are used in design to create detailed profiles that represent specific types of users or groups. These profiles capture essential characteristics without resorting to stereotypes. By including realistic details like behaviours, needs, and demographic information, personas become relatable and useful for understanding different user groups. They help design and development teams build empathy, align their efforts, and develop solutions that are tailored to meet specific needs. Personas are particularly valuable because they help break down traditional marketing segments, leading to more inclusive and effective design. (Kumar, 2013; Stickdorn et al., 2018a, 2018b)

HR Professional Maria's Perspective Bias

When assessing the persona, it's important to consider that her personal views could influence the recommendations and opinions she provides regarding the adoption of the tool. There is ongoing debate about whether AI poses a challenge or risk to traditional recruitment roles. HR professionals might perceive AI recruitment systems as a threat to their jobs, which could result in a reluctance to adopt AI tools for recruitment (Chen, 2023). Other individual biases based on her persona could be for example following:

- Age-related openness or scepticism towards new technology and its impact on traditional roles.
- Influence from being in a tech-forward city like Helsinki, possibly skewing towards innovative solutions.
- Her Guatemalan background might affect her perception of AI's impact on diversity and inclusivity in recruitment.
- Her HR management education may foster a belief in the importance of human judgment, possibly causing reluctance to rely on AI for critical recruitment tasks.
- Living alone, her personal experiences might make her value human interaction more, leading to scepticism about AI's impersonal nature.
- Her work goals emphasize efficiency and personal candidate engagement, which could conflict with her view of AI recruitment.
- Frustrations with high workloads and slow processes might make her open to AI's efficiency, yet the fear of depersonalization and losing touch with candidates could cause hesitation.
- Hobbies like swimming, climbing, and reading indicate a balanced, multifaceted approach to life, which might translate to a desire for a nuanced approach to recruitment.
- Concerns about AI altering HR functions may lead her to question how it aligns with her skill set and the future need for her role.

Recruitment journey



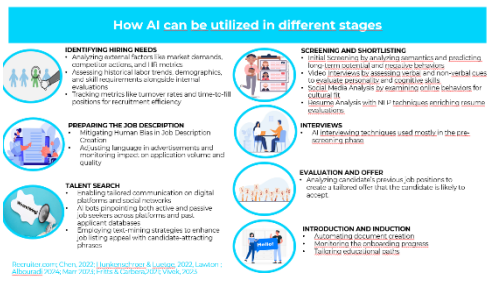
Understanding the recruitment journey

To gain a comprehensive understanding of the recruitment process, the company develops a recruitment journey involving Maria and the recruitment team. Collaborating with line managers, they navigate through each stage of the journey to identify significant frustrations and time-consuming steps. Additionally, they explore the potential integration of AI at each stage to enhance the recruitment journey's efficiency and effectiveness.

Tool: Journey map

A journey map is a visual tool used in design to depict the step-by-step experience a user has with a service. It represents the user's perspective, detailing what happens at each interaction stage, the touchpoints involved, and any challenges or obstacles they might face. Additionally, journey maps often include layers that indicate the user's emotional responses, highlighting positive or negative experiences throughout their journey. These maps can be used to illustrate both current experiences (current-state journey maps) and envisioned future interactions (future-state journey maps). They vary in scope, from broad, high-level overviews of an entire end-to-end experience to highly detailed maps focusing on specific moments. (Kumar, 2013; Stickdorn et al., 2018a).

In this project, the journey map is initially introduced as a current-state depiction, followed by a future-state projection, outlining the capabilities that AI tools may provide for recruitment. AI's usage in different recruitment stages is handled more in details in the chapter "Unveiling AI Solutions: Tackling Recruitment Process Stages and Their Associated Biases".



Stakeholders and their potential biases

Recognizing the personal biases

After a clear problem statement, the company has chosen to proceed with an AI-driven solution. You have been appointed as the lead AI developer responsible for overseeing the implementation of this chosen AI technology.

The company acknowledges the importance of assembling a diverse, multi-stakeholder group to assure a variety of perspectives for the success of the project.

In an effort to address potential unconscious biases within the team, the company has mandated that all members participate in an exercise modeled after Alan Turing called “Positionality Matrix”

Tool: Alan Turing positionality matrix

Reflecting on "positionality" — understanding one's social and cultural background — is crucial in AI development for ethical decision-making. The Alan Turing Institute (2022) underscores the importance of the "positionality matrix," a tool to evaluate how individual traits like age, race, and education influence research and innovation. This self-awareness is key to preventing biases and power imbalances in AI systems. By considering these personal factors, developers can ensure AI solutions are equitable and respect the diverse communities they serve, thus contributing positively without perpetuating injustices.

The Alan Turing positionality matrix is a reflective framework that reveals the impact of individual socioeconomic backgrounds, cultural contexts, and life experiences on judgment and decision-making within a team. It serves as a tool to discern and address the influence of various personal attributes, such as age and ethnicity, on research and development perspectives. This matrix is instrumental in uncovering unconscious biases and navigating power dynamics, thereby fostering a commitment to diversity and equity in AI development. (Alan Turing Institute, 2022)

Taking Stakeholders into Account

On top of the internal stakeholder group taking part in the project, company also needs to consider other related stakeholders.

Therefore, company created a stakeholder map following G.J. Millers definition of the AI development process stakeholder mapping defining the development and usage stakeholders as well as the external stakeholders.

Based on the map, the project manager needs to organize the stakeholders into groups in order to know, how to keep the stakeholders informed and to see what is their influence on the project.

Tool: Stakeholder map

A stakeholder map is a visual tool used to identify and display the roles and relationships of all stakeholders involved in a project. It can range from a simple quadrant showing levels of influence and engagement to a detailed matrix detailing interactions among stakeholders. Stakeholder map helps to clarify stakeholder dynamics and is crucial for understanding and managing network relations. (Giordano et al., 2018; SDT, n.d.)

Tool: Stakeholder analysis matrix

The stakeholder analysis matrix is a visual tool used to categorize and manage stakeholders in a project based on their levels of influence and interest. By arranging stakeholders into a quadrant grid, this matrix helps prioritize engagement strategies. The top-left quadrant represents stakeholders with high influence but low interest, requiring efforts to keep them satisfied without overwhelming them. The top-right quadrant includes highly influential and interested stakeholders who need close management. The bottom-right quadrant, for stakeholders with high interest but lower influence, suggests regular updates to keep them informed. Lastly, the bottom-left quadrant for those with minimal interest and influence requires minimal engagement, focusing on monitoring. (Hoory & Botorff, 2022; Wallbridge, 2023)

Biases in recruitment stages; Workshop

When the stakeholders are clear, company decides to have an ethical workshop going through all the possible biases in different hiring stages.

Stakeholders from the ”Work together” field (previous slide) are invited to join.

What kind of post its would you have filled in?

Tool: Workshop

A workshop is a collaborative design tool where a group of participants engages in interactive activities to address a specific problem or work on a project. Typically facilitated by a leader, workshops can vary in duration from a few hours to several days, depending on the objectives and complexity of the tasks at hand. This method is effective for generating ideas, solving problems, and fostering teamwork and innovation within a group. (Wirtz, 2022)

Unveiling AI Solutions: Tackling Recruitment Process Stages and Their Associated Biases

Algorithms and machine learning technologies are increasingly utilized from the initial stages of the recruitment process. These predictive approaches help anticipate business developments by analysing external factors such as market demands, competitors' moves and shifts in consumer needs. HR metrics play a crucial role in creating these forecasts. They include an analysis of historical labour trends, demographic changes, and evolving skill requirements. Additionally, internal assessments like workload evaluations, managerial observations, employee feedback, and workforce demographics are pivotal in understanding the current state and ongoing evolution of the employee base. Metrics such as turnover rates, time-to-fill positions, and cost-per-hire are crucial for gauging an organization's recruitment efficacy and identifying areas for process improvement. (www.recruiter.com, 2023)

Biases in predictive hiring can arise from various sources, such as the performance level of current employees in specific roles, which may lead to anchoring bias. For example, if an employee excelling in their role needs to be

replaced, analytics might suggest one replacement. However, the high-performance level of the outgoing employee could actually necessitate two replacements. Conversely, the performance of a low-performing employee might inaccurately suggest a need for additional hires to compensate for their lack of productivity. (Chen, 2023, p. 145) Additionally, resistance to change among employees can lead to a status quo bias, where there's reluctance to adapt organizational structures in response to clear business needs. (Sukernek, 2024)

Furthermore, predictive analysis might sometimes overlook critical factors due to a narrow focus. For example, failing to consider broader industry trends or demographic shifts, such as an aging population, can leave a company unprepared for future workforce requirements.

Like other stages in the recruitment process, this phase is also susceptible to feedback loop bias. This type of bias can emerge when AI systems used in recruitment are trained on historical data and inadvertently reinforce their own prejudiced decisions. This occurs in dynamic environments where the AI's decisions influence the data it will learn from in the future. If an AI, already biased by past data, continues to make skewed hiring choices, these decisions feed back into its learning dataset, perpetuating and even amplifying the original biases (Vivek, 2023, p. 0107).

Biases in job description

Artificial intelligence can facilitate creating accurate job descriptions. By tweaking the language of advertisements and monitoring the effect of these modifications on the volume and quality of applications, organizations can enhance the efficiency of their promotional efforts. Moreover, AI can determine which facets of a company, including its culture and achievements, should be highlighted to prospective candidates to elicit the most favourable responses. (Chen, 2023, p. 140)

Job descriptions play a pivotal role in communicating the needs of a company for prospective employees. Yet, biases present in these descriptions can prevent qualified individuals from applying. Such biases often revolve around age, gender, or educational qualifications (Ongig, 2024).

Interestingly, articles suggest that AI has the potential to reduce some of these issues by reviewing job descriptions and identifying language that may contribute to bias. Although AI-generated job descriptions might exhibit fewer biases than human created, the risk of perpetuating existing prejudices remains if the AI tools are not meticulously designed (CareerExperts, 2023). For instance, an anchoring bias may emerge from a manager's previous hiring decisions, where the characteristics and competencies of current top performers influence the creation of new job descriptions (Chen, 2023, p. 145).

Additionally, AI's suggestion to use certain terms in job ads, like “ambitious” or “confident leader,” can unintentionally favour or dissuade specific demographics,

thereby introducing bias (Lawton, 2022). If AI systems are trained using past job descriptions, they risk replicating language and requirements that have historically attracted a less diverse pool of applicants, potentially continuing this trend.

Biases in talent search

Various AI methodologies are employed during the initial engagement phase with prospective candidates. Algorithms facilitate tailored communication across digital platforms and social networks, while AI bots are used in pinpointing both active and passive job seekers across various platforms, or even among a company's database of past applicants. Additionally, AI can deploy text-mining strategies to enhance the appeal of job listings by incorporating phrases that attract the most candidates. Moreover, the implementation of a tone metric by AI can provide recommendations to refine job descriptions, making them more inclusive (Hunkenschroer & Luetge, 2022, p. 992).

AI-driven job advertisements inherently carry the risk of indirect discrimination. This is because employers can specify the audience for their ads using criteria like language skills, educational background, professional experience, or even age and gender, which may prevent certain groups from even seeing these job opportunities, thereby contributing to workplace homogeneity (Chen, 2023, p. 144; Sheard, 2022, p. 624). For instance, Verizon's decision to target a Facebook ad at individuals aged 25 to 36, living in or having visited the US capital, with an interest in finance, might inadvertently exclude other potential job seekers from the ad's reach (Sheard, 2022, p. 624). Algorithms designed to enhance job add efficiency tend to focus on demographics that historically interact more with these advertisements, thus narrowing the diversity of the applicant pool over time (Köchling & Wehner, 2020).

Recruitment biases are further reinforced by recruiters' preferences for specific advertising platforms and their tendency to engage more with candidates from those platforms. This can lead to a lack of diversity since different platforms attract different types of users (Lawton, 2022). Bias in recruitment isn't solely the result of algorithms or employer practices; it also stems from the behaviours of applicants themselves. Relying on social media or other online activities to identify potential candidates can skew the pool towards those who are active online or have the know-how to craft an appealing digital presence. This is particularly evident in highly competitive industries like IT, where the busiest professionals may not regularly update their LinkedIn profiles. Conversely, there are those who might strategically manage their online profiles to enhance their appeal, raising questions about the authenticity of such digitally curated personas. (Naakka, 2018, pp. 46–47)

The issue of transparency is a significant ethical concern in AI-assisted recruitment. The complexity of algorithms can result in a "black box" effect, where the reasoning for decisions remains obscure to applicants (DigitalOcean, n.d.).

Additionally, there are ethical questions regarding the use of candidates' profiles and activities on social media and various platforms for hiring purposes. The standards for handling such data vary internationally, with the European General Data Protection Regulation (GDPR) being among the most stringent (Hunkenschroer & Luetge, 2022, p. 995).

Biases in Screening and shortlisting candidates

AI technologies play a crucial role in the initial screening of applicants by prioritizing those who appear most suitable for the role. Historically, companies relied on scanning resumes for specific keywords. However, modern approaches are more sophisticated, including chatbots and resume parsing tools that search for semantic correlations and assess other qualifications. Some tools even predict a candidate's potential future performance by looking for indicators of long-term commitment or productivity, as well as the absence of negative markers like habitual lateness or disciplinary issues (Hunkenschroer & Luetge, 2022, p. 992).

In the process of candidate shortlisting, AI is also employed in analysing video interviews. AI assistants conduct these interviews, asking a set list of questions and evaluating not just the answers but also analysing the candidate's tone of voice, microfacial expressions, and emotional responses to infer personality traits. Additionally, AI-powered video games are utilized to evaluate traits such as risk-taking behaviour, planning skills, perseverance, and motivation (Hunkenschroer & Luetge, 2022, p. 992).

Further, AI solutions analyse applicants' digital footprints, including social media activities and other online behaviours, through linguistic analysis to gauge their compatibility with the company's culture (Hunkenschroer & Luetge, 2022, p. 992). Beyond social media scrutiny, Neuro-linguistic programming (NLP) techniques are applied at various recruitment stages, including resume analysis, to provide deeper insights (Albaroudi et al., 2024, p. 391).

AI holds the promise of fostering a fairer recruitment environment by reducing unconscious bias. Through an emphasis on predefined abilities and credentials, AI-driven solutions can overlook demographic details like age, gender, and ethnicity, which could unintentionally sway a recruiter's judgement. Such an approach has the potential to enhance workplace diversity, as it ensures that candidates are assessed based on their qualifications and achievements. Furthermore, certain AI applications are specifically developed to assist companies in achieving their diversity and inclusion objectives. They achieve this by examining recruitment trends and proposing adjustments to address any disparities. (DigitalOcean, n.d.)

Despite the potential for screening processes to reduce certain biases, they can inadvertently introduce others if not meticulously managed. Bias may stem from the principles behind model design, the selection of features, and the data used for training (Hunkenschroer & Luetge, 2022, p. 994).

When training data is based on a pool of existing employees or a narrowly defined group that doesn't proportionately represent diverse populations, it can lead to unintentional discrimination against underrepresented groups (Hunkenschroer & Luetge, 2022, p. 994). AI systems, learning from data that may already contain historical biases, have the potential to further amplify these biases. For example, if an AI system is predominantly trained on resumes from a specific demographic, it may unfairly favour candidates from that group (Vivek, 2023). A notable instance occurred in 2018 when Amazon's AI hiring system displayed a preference for patterns dominant among male applicants, thereby discriminating against female candidates. This was because Amazon's training data primarily consisted of resumes from top performers, most of whom were male, leading the algorithm to penalize characteristics typically associated with females (Albaroudi et al., 2024, p. 386; Hunkenschroer & Luetge, 2022, p. 994).

Furthermore, training bias can also emerge when an algorithmic tool fails to adequately consider all skills and traits relevant to the job. If the tool overly focuses on technical abilities without accounting for interpersonal skills, crucial competencies like communication might be overlooked. Aggregation bias, where false generalizations are made about entire populations, can also occur, leading to the exclusion of individuals based on assumptions about group characteristics. For instance, algorithms might presume younger people to have superior technological skills, thus overlooking older candidates who are equally qualified (Albaroudi et al., 2024, p. 386).

The assessment of a candidate’s digital footprint as part of their screening can also perpetuate biases, reflecting the varying perceptions of what should be presented on social media and the limitations people face in updating their profiles. A study by North Carolina State University involving interviews with 61 HR professionals revealed a tendency to value personal interests, such as hiking or celebrating Christmas, over professional traits. This approach benefits certain demographics over others and assumes characteristics like being "energetic" or "active" are inherently superior, potentially discriminating against older or disabled individuals. Introducing AI into this process risks embedding these prejudices into the algorithms (Shipman, 2021).

The use of emotional recognition software in video assessments must carefully consider the diverse intonations across languages and the potential for systematic disadvantage to specific races or ethnic groups. Moreover, some individuals may find it challenging to behave naturally in front of a camera, leading to inaccurate assessments by the tool (Hunkenschroer & Luetge, 2022, p. 995).

Privacy and informed consent in using candidate information for recruiting are critical ethical considerations, with regulations varying internationally. The European General Data Protection Regulation (GDPR) is among the strictest frameworks, designed to safeguard EU citizens' rights by regulating the collection, storage, and processing of personal data and requiring informed consent for any personal data processing activities. Nevertheless, the dilemma

arises when considering whether applicants would risk opting out, fearing it might jeopardize their job prospects (Hunkenschroer & Luetge, 2022, p. 995).

Biases in interviewing the candidates

One reason for using the AI based interviewing techniques can be the language barriers in multinational recruitment, necessitating local talent acquisition and multilingual HR professionals. AI mitigates this by enabling the interview of diverse language speakers without the need for HR to know every local dialect, thanks to advancements in natural language processing (NLP). This technology allows businesses to efficiently bridge the language gap in global hiring without relying on translators. (Albaroudi et al., 2024, p. 392)

AI-based interviewing tools are, however, more often used to automate the first round of interviews and thereby contributing to the evaluation of candidates (Fritts & Cabrera, 2021, p. 792). Therefore, AI based interviewing solutions and the possible biases they are creating, are handled in the previous chapter “Screening and shortlisting the candidates.”

Vivek (2023) suggests that due to nuanced understanding and emotional intelligence humans possess, the final decision-making power should remain with human recruiters. Fritts & Cabrera (2021) mention also that the concerns around AI-driven processes for identifying, screening, interviewing, and onboarding candidates, might lead to dehumanization by removing personal interactions. This could be ethically problematic as the artificial values embedded in hiring algorithms might undermine the employee-employer relationship by stripping away the inherently human aspects of traditional interactions.

Biases in evaluation and offer creation

After choosing a candidate, the company must extend a job offer. AI algorithms can evaluate candidate's past positions to formulate an offer that the candidate is probable to accept. However, these tools may exacerbate existing inequalities in initial and ongoing salaries among different genders, races, and other demographics. This happens because these tools often rely on historical wage data, which may contain inherent biases. (Lawton, 2022)

Biases in onboarding

Implementing AI in onboarding can reduce routine tasks of HR professionals by automating document management, scheduling communications, and using chatbots for initial feedback and queries. It also assesses new employees' progress and ensures compliance with training requirements while providing analytics on onboarding trends. (Marr, 2023)

AI interfaces can tailor educational paths to fit individual skills and learning styles, ensuring that new hires receive the support they need for success. HR departments can utilize insights from automated interactions and feedback to understand new hires' capabilities and growth areas. Predictive analytics can anticipate the needs of new employees, offering proactive resources and support. (Marr, 2023). AI can also be used in matching the newcomer with the work mentor based on their skills, interests and expertise. (Featured, 2023)

As with earlier phases, biases in AI-driven onboarding may arise from the use of training data rooted in historical records that carry previous prejudices. Furthermore, the algorithms might inadvertently embody the developers' subconscious biases in shaping educational pathways.

6. Screening and shortlisting the candidates: dataset definition

Screening and shortlisting the Candidates; dataset definition

While the company is looking into all aspects and possibilities in AI recruiting, currently the company is intensively evaluating AI on the screening and shortlisting phase. This involves an in-depth analysis of resume scanning and the alignment of candidates to suitable job vacancies.

By deploying AI tools to automate the resume screening process, the goal is to quickly and precisely pinpoint candidates who fulfill the job requirements. The adoption of such technology is expected to lead to a more strategic use of HR resources, enhance the caliber of candidates making it to the shortlist, and, by aligning candidates' skills and backgrounds with the company's requirements, increase the overall effectiveness of the hiring process.

First, the company needs to define the datasets, which are the basis of the recruiting system and from which the algorithms take the data from.

Training AI

Training datasets play a crucial role in the development of algorithms. These datasets are the foundation on which algorithms are built. The purpose of these datasets is to serve as the "ground truth" or factual basis that instructs the algorithm on how to interpret and process large quantities of data. They teach the algorithm to understand the relationships between input variables (the information fed into the system) and an output variable (the prediction or decision the system makes based on the inputs). Significant challenge arises when the training data itself contains biases. These biases can be a reflection of historical inequalities, prejudices, or simply the result of unrepresentative or incomplete data collection. When an algorithm is trained on biased data, it learns these biases and perpetuates them in its predictions and decisions. This phenomenon is commonly described as the "bias in, bias out" problem. It means that if the input data (bias in) is biased, the output of the machine learning model (bias out) will also be biased. This can lead to unfair, discriminatory, or flawed outcome affecting which candidates are selected or rejected based on biased criteria rather than their true qualifications or potential. (Sheard, 2022)

Training AI on company's internal data

When training AI applications for CV screening, companies have access to a variety of internal data sources that can be instrumental in the learning process. (Fernandes, 2021; Ma, 2024; Sheard, 2022)

One such source is the employee performance data coupled with historical employment data, which encompasses internal records of employee assessments and past hiring outcomes. Additionally, companies can utilize the accumulation of CVs submitted over time, which includes a wealth of historical CVs and resumes that were received directly for job applications.

Another valuable resource is the 'false negative cases' data, which refers to information on candidates who were mistakenly rejected in previous recruitment cycles. Analysing these cases can provide insights into patterns that the AI should avoid repeating. Moreover, companies can extract specific keywords and phrases that are frequently encountered in the CVs collected, using them to refine the AI's ability to identify relevant qualifications and experience.

Lastly, educational background information, which is a standard component of candidate applications, can also be leveraged to enhance the AI's understanding of academic qualifications and their relevance to job roles.

Training AI on company's external data

Possible sources for external data in AI training have been suggested and include several platforms and methods for gathering information. (Banerjee, 2022) Social

media platforms are a rich source where individuals share personal interests, experiences, and professional achievements, providing insights into a candidate's personality, skills, and professional network. (Albaroudi et al., 2024; Albassam, 2023; Chaker, 2018; Filtered, n.d.; Heymans, 2022; Shipman, 2021) Additionally, professional websites designed for networking and career development, such as LinkedIn, contain detailed profiles that include work history, education, skills, endorsements, and professional accomplishments. These profiles offer a comprehensive view of potential talents. Finally, internet searches conducted by companies can unearth publicly available information about a candidate from sources like news articles, personal blogs, publications, or other web presences. This broad array of information provides valuable context about a person's qualifications, achievements, and character. (Chaker, 2018)

7. Developing the algorithmic model

Job advertisement; key word selection for algorithm

The company has recently published a job description and received hundreds of applications in response.

AI screening and shortlisting prototype will scan applicants' CVs for keywords and other information believed to be relevant to successful hires, such as experience, job titles, previous employers, universities and degrees. Based on this, the system creates a structured applicant profile, and all candidates are scored and ranked. (Sheard, 2022)

AI significantly streamlines the applicant screening process, which is traditionally one of the most time-intensive steps in recruitment. During the 'screening' stage of the recruitment funnel, the evaluation of job candidates based on their experience, skills, and other pertinent characteristics is crucial for creating an interview shortlist. AI can be used for scanning CVs for keywords and other data associated with successful hires—such as job titles, previous employers, educational institutions, and qualifications—thereby streamlining the process of identifying suitable candidates. (Dennis, 2023; Fernandes, 2021; Sheard, 2022)

By utilizing AI tools, recruiters can quickly identify and prioritize top candidates, thereby enhancing efficiency and improving key recruitment metrics such as time-to-hire. AI systems automatically screen out candidates who fail to meet

specific pre-set criteria, parse resumes to highlight relevant experience, and rank applicants according to their suitability for the role. This automation allows recruiters to focus more on evaluating candidates at later stages of the recruitment funnel. Additionally, AI can scan resumes during the initial resume screening phase, pinpointing candidates who meet specific criteria. Machine learning algorithms then assess these candidates' skills and qualifications against the job specifications to compile a list of suitable applicants. (ResumeMent, 2023; Dennis, 2023)

The job description provides essential details about the role, including required skills, qualifications, and experience, which the AI uses to screen the CVs. Without a detailed job description, the AI would lack the necessary parameters for what to look for in the CVs.

Job description for Communication Assistant
We are looking for a communications assistant to be responsible for the creation of content such as media releases, blogs, and social media posts on behalf of our company. You will also be monitoring media and campaign coverage and attending internal and external events.

Responsibilities
- Develop and manage diverse content, including media releases, blogs, and social media posts. aligned with the company's strategic goals; monitor media and campaign coverage.
- Support the implementation of internal and external communications strategies and assist in managing the company's image.
- Organize marketing events and provide comprehensive administrative support. including maintaining event calendars and updating contact lists.
- Compile and prepare presentations and reports while tracking project progress and media exposure.

Requirements
- Holds a Bachelor's degree in communications. marketing, or related field, with excellent verbal and written communication abilities.
- Proficient in social media strategies and media relations, with a creative and innovative approach.
- Strong organizational skills and attention to detail, capable of multitasking and maintaining positive interpersonal relationships.
- Skilled in using office management and design software, such as Photoshop and InDesign, and knowledgeable in various social media platforms.

Advancements in AI have moved beyond simple keyword matching. Initially, companies used algorithms to scan resumes for preselected keywords or phrases. Today's AI technologies, including chatbots and sophisticated resume-parsing tools, now assess semantic relevance and related terms to more accurately determine a candidate's qualifications. Additionally, machine learning algorithms are employed to predict future job performance based on various indicators, such as work tenure and productivity. These tools can also suggest the most suitable job openings for candidates based on their profiles. (Hunkenschroer & Luetge, 2022)

However, in this course, we use simple keyword matching as an example to illustrate how an AI CV screening application processes applications.

List of Specific Keywords Used for Candidate Screening

"Bachelor's degree in communications"
"Bachelor's degree in marketing" "Content creation"
"Media releases"
"Blogs"
"Social media posts"
"Monitoring media" and "campaign coverage"
"Supporting implementation of communications strategies"
"Managing company's image"
"Organizing marketing events"
"Providing comprehensive administrative support" "Updating contact lists"
"Compiling presentations" and "reports" "Tracking project progress" and "media exposure"
"Proficient in social media strategies" and "media relations"
"Creative and innovative approach"
"Strong organizational skills" and "attention to detail"
"Capable of multitasking"
"Maintaining positive interpersonal relationships" "Skilled in using office management" and "design software"
"Photoshop" and "InDesign"

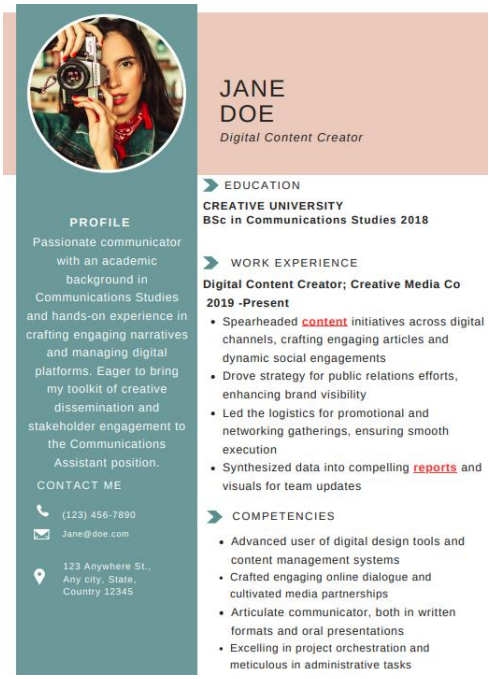
According to Sheard (2022), resumes are typically reviewed by human eyes only if they align with the job advertisement criteria, while the rest may be disregarded, highlighting a potential concern where a significant portion of resumes might be prematurely excluded.

Next, we will focus on analysing two applications to understand how the AI application prototype evaluates candidates during the screening phase. This analysis will help to identify if there are any biases influencing the selection process.

Comparing two applicant's CVs towards the key word phrasing

The company organizes a workshop to explore the details of AI-driven CV screening, using a case study from their own operations.

During the workshop, they analyze two CVs side-by-side to uncover how differences in wording and phrasing can inadvertently lead to overlooking qualified candidates.



This example illustrates the impact of wording and phrasing differences in two CVs and how they can affect the outcome of an AI screening process. Despite possessing relevant qualifications, the second CV may be overlooked due to its descriptions not aligning closely with the specific keywords established by the company for screening. This demonstrates the importance of matching the language in a CV to the keywords expected by an AI system.

Aspect	CV1	CV 2
Job Titles and Terminology	Uses standard terms like "Communications Assistant"	Uses titles like "Digital Content Creator"
Academic Background	Explicitly mentions "Bachelor's Degree in Marketing"	States "BSc in Communications Studies"
Describing Tasks and Skills	Direct match with keywords like "Content creation", "Media releases"	Uses varied language like "crafting engaging narratives", "spearheaded content initiatives"
Technical and Software Skills	Specifies "Photoshop" and "InDesign"	General mention of "digital design tools and content management systems"
Direct Keyword Matches	Closely matches many specified keywords	Uses different phrasing that may not match the specific Keywords set for screening

8. How to measure fairness in machine learning solutions

Testing the fairness of the solution

Evaluating the algorithm prototype revealed inconsistencies in CV analysis, prompting refinements for a more comprehensive understanding of CV attributes. The model is now prepped for a testing phase prior to real-world deployment.

While several aspects require assessment, the focus here will be on testing the application's fairness to determine the extent of any biases incurred during the process. Fairness metrics are employed to measure the level of bias present in the application's decisions.

To create systems that are equitable for all and free from biases like gender or racial discrimination, it's important to evaluate the solution's performance across diverse population segments. While there are various methods to assess a solution's fairness, in this context, we will focus exclusively on binary classification tests. (*Machine Learning and Fairness*, 2021)

Fairness has emerged as a crucial topic in machine learning due to its increasing application across various fields. Extensive research has focused on this area because machine learning systems heavily depend on data, which, when sourced

from humans, can be inherently biased. Defining fairness in mathematical terms has proven challenging, leading to a lack of consensus on standard formulations. However, as noted by Zhong (2018), most definitions of fairness coalesce around several key concepts:

- Unawareness
- Demographic parity
- Equalized odds
- Predictive rate parity
- Individual fairness
- Counterfactual fairness

Demographic Parity, Equalized Odds, and Predictive Rate Parity are typically grouped under the umbrella of "group fairness." (Zhong, 2018). This concept will be explored further in this section, particularly through the lens of AI in recruiting.

On top of group fairness Castelnovo et al. (2022) divide the fairness categories under individual fairness and Causality based fairness, including also the topics mentioned by the definitions of Zhong (2018).

Confusion matrix

One of the binary classification methods, is the confusion matrix (*Machine Learning and Fairness*, 2021). To grasp fairness metrics, we start with the concept of a confusion matrix. This tool summarizes the predictions made by a model in comparison to the actual outcomes, or ground truth, on which it was trained. The confusion matrix displays the counts of both correct and incorrect predictions, providing a clear view of the model's performance and its explainable validity. (Saplicki, 2022)

	Unqualified	Qualified
Reject	True Negative (TN)	False Negative (FN)
Hire	False Positive (FP)	True Positive (TP)

IMAGE SOURCE | Picture 1 - Confusion matrix (Modified from (Machine Learning and Fairness, 2021)

A confusion matrix helps us categorize the outcomes of the hiring process. Qualified candidates who are successfully hired are deemed true positives (TP) and unqualified candidates appropriately rejected are true negatives (TN). Conversely, false positives (FP) refer to unqualified candidates who are mistakenly hired, while false negatives (FN) are qualified candidates wrongly rejected. A confusion matrix simply categorizes applicants into distinct groups based on their classification outcomes. The four values produced by a confusion matrix are utilized to calculate standard machine learning metrics, including accuracy and precision, across different demographic groups. (*Machine Learning and Fairness*, 2021; Teodorescu, 2020)

In the recruitment scenario being discussed, we will focus specifically on gender demographics. Although it is important to consider other demographic factors, this example will solely concentrate on gender.

Let's assume there is an equal number of 100 male and female applicants. The system has accurately identified 75 individuals from each group as either qualified or unqualified, demonstrating that the accuracy rate is consistent across both demographics. (*Machine Learning and Fairness*, 2021)

MEN	Unqualified	Qualified	WOMEN	Unqualified	Qualified
Reject	**15 (TN)**	5 (FN)	Reject	**60 (TN)**	20 (FN)
Hire	20 (FP)	**60 (TP)**	Hire	5 (FP)	**15 (TP)**

IMAGE SOURCE | Picture 2 - Confusion matrix example from recruiting (Modified from Machine Learning and Fairness, 2021)

Demographic Parity

Building on the provided data, we'll examine how various fairness metrics apply in this context, beginning with demographic parity. This metric considers only the classifier's output, without regard to the actual labels. Demographic parity is one of the suggested metrics, when evaluating an automated hiring system. (*Machine Learning and Fairness*, 2021)

Demographic Parity, commonly known as Independence or Statistical Parity, is a criterion for fairness where the probability of a favourable outcome, such as being hired, is not influenced by a protected characteristic like gender (Zhong, 2018).

MEN	Unqualified	Qualified
Reject	15 (TN)	5 (FN)
Hire	**20 (FP)**	**60 (TP)**

WOMEN	Unqualified	Qualified
Reject	60 (TN)	20 (FN)
Hire	**5 (FP)**	**15 (TP)**

IMAGE SOURCE | Picture 3 - Demographic parity example from recruiting (Modified from Machine Learning and Fairness, 2021)

In the example scenario of recruiting (picture 3), if 80 out of 100 male applicants and only 20 out of 100 female applicants are hired (picture 3) demographic parity is not met. However, this might be acceptable if the male applicants are indeed more qualified than the female applicants, like it seems to be in this case. (*Machine Learning and Fairness*, 2021).

However, demographic parity is challenging to achieve when individuals belong to multiple protected groups, as equal probabilities might not be feasible across

all demographics. Demographic parity can sometimes inadvertently favour less qualified individuals due to its group-level focus, potentially leading to individual-level unfairness. For example, increasing the proportion of less qualified female applicants might decrease the chances of hiring qualified male applicants. (Teodorescu, 2020) Therefore, this measure solely considers the outcome, not the true labels, meaning it doesn't account for the actual qualifications of applicants. It holds that the ratio of hired to total applicants should be similar across all. (*Machine Learning and Fairness*, 2021).

Predictive parity

Sufficiency, as viewed from the perspective of those who receive the same decision from a model, mandates equal error rates regardless of sensitive attributes. This concept is encompassed by Predictive Parity, also known as the outcome test. It ensures that error rates are balanced for individuals grouped by the decisions they receive, rather than by actual outcomes. Unlike demographic parity, predictive parity considers both the classifier's decisions and the true outcomes. It assesses whether the likelihood of an applicant being suitable for a position is consistent across different groups, provided the AI has selected them for hiring. The expectation is that this probability should be nearly identical for all groups considered. (Castelnovo et al., 2022; *Machine Learning and Fairness*, 2021).

MEN	Unqualified	Qualified
Reject	15	5
Hire	20	**60**

WOMEN	Unqualified	Qualified
Reject	60	20
Hire	5	**15**

IMAGE SOURCE | Picture 4 - Predictive parity example from recruiting (Modified from *Machine Learning and Fairness*, 2021)

In our recruiting example the system meets the criteria for predictive parity as three-quarters of the hired applicants from both groups are indeed qualified (*Machine Learning and Fairness*, 2021).

False positive rate balance, False negative rate balance & Equalized odds

The metric known as the **false positive rate** asserts that the likelihood of an unqualified candidate being erroneously hired should be consistent across all demographics. In simpler terms, the system should equally misjudge unqualified candidates, irrespective of their gender. (*Machine Learning and Fairness*, 2021)

False positive rate

MEN	Unqualified	Qualified
Reject	15	5
Hire	**20**	60

WOMEN	Unqualified	Qualified
Reject	60	20
Hire	**5**	15

IMAGE SOURCE | Picture 5 - False positive rate example from recruiting (Modified from *Machine Learning and Fairness*, 2021)

In this example (picture 5) the false positive rates are different with a substantially larger fraction of unqualified men being hired than unqualified women.

False negative rate balance is essentially the same thing as the false positive rate, but for false negative rates. The concept of false negative rate parity suggests that the likelihood of erroneously rejecting qualified candidates should be uniform across all groups.

False negative rate

MEN	Unqualified	Qualified
Reject	15	**5**
Hire	20	60

WOMEN	Unqualified	Qualified
Reject	60	**20**
Hire	5	15

IMAGE SOURCE | Picture 6. False negative rate example from recruiting (Modified from *Machine Learning and Fairness*, 2021)

In this instance (picture 6) there is a noticeable discrepancy in false negative rates, with qualified female candidates facing rejection at a higher rate than their male counterparts.

Equalized odds combines these two attributes by satisfying both false positive rate balance and false negative rate balance, meaning that the system must handle both types of errors—incorrectly accepted unqualified candidates and incorrectly rejected qualified candidates—equitably across different groups, ensuring all similarly qualified candidates receive consistent treatment. (*Machine Learning and Fairness*, 2021). Equalized odds is hard to achieve, but when achieved it can be considered as one the highest levels of algorithmic fairness. (Teodorescu, 2020)

Trade-offs with metrics

A system cannot simultaneously fulfil predictive parity, false positive rate balance, and false negative rate balance due to mathematical constraints. If a system meets two of these criteria, it will inevitably fall short on the third. (*Machine Learning and Fairness*, 2021)

This theorem underscores the inherent trade-offs between different fairness goals. As we strive for fairness, we often encounter conflicting objectives, and achieving perfect balance across all dimensions remains elusive.

In practice, decision-makers must carefully weigh these trade-offs, consider the context, and make informed choices based on the specific needs of the application. While perfection may be unattainable, continuous efforts toward fairness and transparency are essential in algorithmic decision-making systems.

Finally, it's important to acknowledge that numerous facets of fairness elude quantification due to their sociotechnical complexity. Consequently, not all aspects of fairness can be encapsulated by metrics. (*Machine Learning and Fairness*, 2021)

Individual fairness

Individual fairness differs from group-based fairness (the previously discussed metrics) criteria by focusing on the treatment of individuals. This concept asserts that individuals with similar characteristics should receive comparable outcomes. Determining a metric to measure individual similarity is complex. Consider three job candidates: A with a bachelor's degree and one year of related experience; B with a master's degree and the same amount of experience; and C with a master's degree but no experience. Deciding whether A is more like B or C, and quantifying that similarity, is challenging without being able to compare their job performance, which isn't possible if all three are not hired. (Zhong, 2018)

9. Model deployment

Informing applicants about using AI in recruting

Following careful internal evaluations, the company is set to launch its screening tool selectively on open roles to gather initial user and applicant feedback.

It's crucial for applicants to be continuously aware that they're engaging with an AI-driven recruitment system, fostering trust. Awareness about the advantages of AI tools should be established upfront, enhancing applicants' willingness to use interactive technology like chatbots. Additionally, refining AI recruitment processes and ensuring clarity about the system's workings are key to its adoption and success among candidates.

Also, it should be made sure that there is a human overseeing and intervening when needed. (Chen, 2023)

10. Operation and monitoring

After deployment

While internal assessments of fairness were conducted before launch, the company must also be ready for independent external audits.

It’s also advantageous to continuously seek feedback post-deployment and establish ongoing engagement channels for stakeholders. This involves keeping workers informed and involved in consultations and participative processes during the entire AI system implementation within the organization. (HLEG Trustworthy AI)

Auditing AI

While internal Fairness measurements are done prior the deployment, an external audit is also something that the company needs to be prepared for.

According to Ahmed (2020) auditors evaluating AI applications should prioritize two aspects: ensuring compliance with data subject rights and assessing technological risks in machine learning and cybersecurity. The audit begins with defining its scope, objectives, and the AI-related risks to the organization, typically documented in a risk and control matrix within established frameworks.

Key risks involve the misalignment of IT strategies with business objectives, ineffective governance, and accountability structures. Compliance-wise, auditors must grasp data privacy principles and their implications on individuals' rights under regulations such as the GDPR. (Ahmed, 2020)

There exist several guidelines for auditing AI. As AI reshapes business and society, auditors need to ensure readiness to tackle emerging challenges and verify the effectiveness of controls and governance around AI. (Ahmed, 2020)

11. Conclusion

This course unit focused on examining algorithmic biases in a practical context. The theme was explored through a hypothetical scenario in which an international company develops an AI solution for recruiting. As AI technologies are integrated into recruitment, they offer substantial benefits by enhancing efficiency and potentially increasing workplace diversity by emphasizing qualifications over demographic characteristics. However, the implementation of these technologies also raises significant ethical challenges, particularly in identifying and mitigating algorithmic biases.

AI systems can inadvertently perpetuate existing biases found in their training data. These biases might reflect historical disparities or societal inequalities, leading to discriminatory practices in recruitment processes. Addressing these biases is not only an ethical imperative but also a legal necessity, as the regulatory landscape surrounding AI in recruitment is continuously evolving. Organizations must remain vigilant and proactive in adhering to these standards to prevent discrimination.

To effectively mitigate these challenges, it is crucial to use sophisticated tools specifically designed for identifying and correcting biases in AI systems. Regular monitoring and independent audits are essential, as they help organizations identify emergent biases and adjust their AI systems accordingly.

Moreover, maintaining a balance between AI-driven processes and human oversight is critical. While AI can significantly streamline recruitment processes, the human element remains indispensable in interpreting contexts that AI might misjudge. Human oversight ensures that recruitment remains a sensitive and inclusive process, respecting the nuances of human experiences and values.

In summary, while AI in recruitment offers considerable advantages, leveraging this technology responsibly is crucial. Organizations should focus on continuously identifying biases and employing advanced tools to mitigate them. This balanced approach ensures that recruitment processes uphold the highest standards of fairness and ethical practice, enabling companies to embrace technological advancements while safeguarding diversity and equity in the workplace.

12. References

Ahmed, H. S. A. (2020, December 21). *Auditing Guidelines for Artificial Intelligence*. ISACA. https://www.isaca.org/resources/news-and-trends/newsletters/atisaca/2020/volume-26/auditing-guidelines-for-artificial-intelligence

Alan Turing Institute. (2022). *Reflect on Purpose, Positionality, and Power—Turing Commons*. https://alan-turing-institute.github.io/turing-commons/skills-tracks/aeg/chapter5/reflect/

Albaroudi, E., Mansouri, T., & Alameer, A. (2024). A Comprehensive Review of AI Techniques for Addressing Algorithmic Bias in Job Hiring. *AI*, *5*(1), 383–404. https://doi.org/10.3390/ai5010019

Albassam, W. A. (2023). The Power of Artificial Intelligence in Recruitment: An Analytical Review of Current AI-Based Recruitment Strategies. *International Journal of Professional Business Review*, *8*(6), e02089. https://doi.org/10.26668/businessreview/2023.v8i6.2089

ALTAI. (2020). *The Assessment List for Trustworthy Artificial Intelligence*. https://altai.insight-centre.org/

Arivu Recruitment and Consulting. (2023, July 13). *Pros and Cons of Using Artificial Intelligence (AI) in Recruiting | LinkedIn*. https://www.linkedin.com/pulse/pros-cons-using-artificial-intelligence/

Banerjee, S. (2022, February 20). *Big data in Recruitment Sector | LinkedIn*. https://www.linkedin.com/pulse/big-data-recruitment-sector-shantanu-banerjee-phd/

Bursell, M., & Roumbanis, L. (2024). After the algorithms: A study of meta-algorithmic judgments and diversity in the hiring process at a large multisite company. *Big Data & Society*, *11*(1), 20539517231221758. https://doi.org/10.1177/20539517231221758

CareerExperts. (2023, April 24). Reducing Bias in Hiring with AI-Powered Job Description Generation. *Career Experts*. https://www.careerexperts.co.uk/management-leadership/ai-powered-job-description

Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., & Cosentini, A. C. (2022). A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, *12*(1), 4209. https://doi.org/10.1038/s41598-022-07939-1

Chaker, N. (2018, October 24). *Candidate Data Sources: The Recruiter's Checklist*. https://beamery.com/resources/blogs/candidate-data-sources-the-recruiter-s-checklist

Chen, Z. (2023). Collaboration among recruiters and artificial intelligence: Removing human prejudices in employment. *Cognition, Technology & Work*, *25*(1), 135–149. https://doi.org/10.1007/s10111-022-00716-0

Dennis, J. (2023, May 22). *AI Recruiting: Uses, Advantages, & Disadvantages 2024*. TechnologyAdvice. https://technologyadvice.com/blog/human-resources/ai-recruiting/

DigitalOcean. (n.d.). *How to Use AI in Hiring: Techniques and Tools*. Retrieved April 26, 2024, from https://www.digitalocean.com/resources/article/ai-in-hiring

Featured. (2023, December 22). *How these 6 leaders are using AI-powered onboarding*. Fast Company. https://www.fastcompany.com/90996367/how-leaders-using-ai-powered-onboarding

Fernandes, R. F. (2021). Big Data as a Tool to Enhance Recruitment Processes. *E-Journal of International and Comparative LABOUR STUDIES*, *10*(03).

Filtered. (n.d.). *What is AI Resume Screening and How Can it Benefit Your Enterprise? | Filtered Blog*. Retrieved March 14, 2024, from https://www.filtered.ai/blog/ai-resume-screening-fd

Friedman, B., & Nissenbaum, H. (1996). Bias in Computer Systems. *ACM Transactions on Information Systems*, *14*(3), 330–347.

Fritts, M., & Cabrera, F. (2021). AI recruitment algorithms and the dehumanization problem. *Ethics and Information Technology*, *23*(4), 791–801. https://doi.org/10.1007/s10676-021-09615-w

Giordano, F., Morelli, N., Götzen, A. D., & Hunziker, J. (2018, June). *The stakeholder map: A conversation tool for designing people-led public services*. ServDes2018 - Service Design Proof of Concept, Politecnico di Milano.

Google for Developers. (2022). *Fairness: Types of Bias | Machine Learning*. Google for Developers. https://developers.google.com/machine-learning/crash-course/fairness/types-of-bias

Harvard John A. Paulson School of Engineering and Applied Sciences. (2023, December 6). *How Can Bias Be Removed from Artificial Intelligence-Powered Hiring Platforms?* https://seas.harvard.edu/news/2023/06/how-can-bias-be-removed-artificial-intelligence-powered-hiring-platforms

Harwell, D. (2019a, July 7). FBI, ICE find state driver's license photos are a gold mine for facial-recognition searches. *Washington Post*. https://www.washingtonpost.com/technology/2019/07/07/fbi-ice-find-state-drivers-license-photos-are-gold-mine-facial-recognition-searches/

Harwell, D. (2019b, December 19). Federal study confirms racial bias of many facial-recognition systems, casts doubt on their expanding use. *Washington Post*. https://www.washingtonpost.com/technology/2019/12/19/federal-study-confirms-racial-bias-many-facial-recognition-systems-casts-doubt-their-expanding-use/

Heymans, Y. (2022, September 7). *Machine learning in recruitment: A deep dive*. https://www.herohunt.ai/blog/machine-learning-in-recruitment-a-deep-dive

Hidalgo, M. A. S. (2019). *Design of an Ethical Toolkit for the Development of AI Applications* [Delft University of Technology]. http://resolver.tudelft.nl/uuid:b5679758-343d-4437-b202-86b3c5cef6aa

Hoory, L., & Botorff, C. (2022, August 7). *What Is A Stakeholder Analysis? Everything You Need To Know – Forbes Advisor*. https://www.forbes.com/advisor/business/what-is-stakeholder-analysis/

Hunkenschroer, A. L., & Luetge, C. (2022). Ethics of AI-Enabled Recruiting and Selection: A Review and Research Agenda. *Journal of Business Ethics*, *178*(4), 977–1007. https://doi.org/10.1007/s10551-022-05049-6

Javaid, S. (2024, January 12). *Facial Recognition: Best Practices & Use Cases in 2024*. AIMultiple: High Tech Use Cases & Tools to Grow Your Business. https://research.aimultiple.com/facial-recognition/

Köchling, A., & Wehner, M. C. (2020). Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research*, *13*(3), 795–848. https://doi.org/10.1007/s40685-020-00134-w

Kumar, V. (2013). *101 Design Methods—A Structured Approach for Driving Innovation in Your Organization*. John Wiley & Sons, Inc. https://learning.oreilly.com/library/view/101-design-methods/9781118083468/cvi.xhtml

Lange, A. R., & Duarte, N. (2018, April 4). Understanding Bias in Algorithmic Design. *ASME ISHOW / IDEA LAB*. https://medium.com/impact-engineered/understanding-bias-in-algorithmic-design-db9847103b6e

Lawton, G. (2022, August 29). *AI hiring bias: Everything you need to know | TechTarget*. HR Software. https://www.techtarget.com/searchhrsoftware/tip/AI-hiring-bias-Everything-you-need-to-know

Lee, N. T., Resnick, P., & Barton, G. (2019, May 22). *Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms | Brookings*. https://www.brookings.edu/articles/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Ma, Y. (2024). Intelligent Recommendation Method for Talent Resources Based on Big Data. *Intelligent Computing Technology and Automation*. https://doi.org/10.3233/ATDE231176

Machine Learning and Fairness. (2021, May 24). Microsoft Research webinars and tutorials. https://youtu.be/ZtN6Qx4KddY?si=kWo9MxkNQ7ko1HUB

Marr, B. (2023, December 12). *AI-Enhanced Employee Onboarding: A New Era In HR Practices*. Forbes. https://www.forbes.com/sites/bernardmarr/2023/12/12/ai-enhanced-employee-onboarding-a-new-era-in-hr-practices/

Naakka, S.-S. (2018). *BIG DATA KILPAILUETUNA SUORAHAKU- KONSULTOINNIN EHDOKASHAUISSA Enemmän, nopeammin ja parempia IT-osaajia big datalla?* [University of Turku]. https://www.utupub.fi/bitstream/handle/10024/145309/Naakka%20Sara-Selia.pdf?sequence=1&isAllowed=y

Ongig. (2024). *Job Description Software | Ongig*. https://www.ongig.com

ResumeMent. (2023, October 16). *From Resumes to Algorithms: The Role of AI in Screening Candidates | LinkedIn*. https://www.linkedin.com/pulse/from-resumes-algorithms-role-ai-screening-candidates-resumement/

Saplicki, C. (2022, November 11). Fairness Explained: Definitions and Metrics. *Medium*. https://medium.com/ibm-data-ai/fairness-explained-definitions-and-metrics-9690f8e0a4ea

SDT. (n.d.). *Tools | Service Design Tools*. Retrieved April 18, 2024, from https://servicedesigntools.org/tools.html

Sheard, N. (2022). Employment Discrimination by Algorithm: Can Anyone Be Held Accountable? *University of New South Wales Law Journal*, *45*(2). https://doi.org/10.53637/XTQY4027

Shipman, M. (2021, March 4). Social media checks can bring bias into hiring. *Futurity*. https://www.futurity.org/cybervetting-human-resources-bias-2527382-2/

Stickdorn, M., Hormess, M. E., Lawrence, A., & Schneider, J. (2018a). *This is Service Design Doing*. O'Reilly Media, Inc. https://learning.oreilly.com/library/view/this-is-service/9781491927175/ch03.html

Stickdorn, M., Hormess, M., Lawrence, A., & Schneider, J. (2018b, July). *#TISDD Method Library*. This Is Service Design Doing. https://www.thisisservicedesigndoing.com/methods

Sukernek, W. (2024). *How to Spot Bias in Hiring*. FloCareer. https://blog.flocareer.com/how-to-spot-bias-in-hiring

Teodorescu, M. (2020). *Fairness Criteria | Exploring Fairness in Machine Learning for International Development | Supplemental Resources*. MIT OpenCourseWare. https://ocw.mit.edu/courses/res-ec-001-exploring-fairness-in-machine-learning-for-international-development-spring-2020/pages/module-three-framework/fairness-criteria/

UNESCO. (2023). *Ethical impact assessment. A tool of the Recommendation on the Ethics of Artificial Intelligence*. UNESCO. https://doi.org/10.54678/YTSA7796

Vivek, R. (2023). Enhancing diversity and reducing bias in recruitment through AI: A review of strategies and challenges. *Информатика. Экономика. Управление - Informatics. Economics. Management*, *2*(4), 0101–0118. https://doi.org/10.47813/2782-5280-2023-2-4-0101-0118

Wallbridge, A. (2023, April 13). *How To Conduct A Bullet-Proof Stakeholder Analysis Matrix In 4 Useful Steps*. TSW Training. https://www.tsw.co.uk/blog/leadership-and-management/stakeholder-analysis-matrix/

Wirtz, D. (2022, August 3). *What Is a Workshop? (+2 Examples) | Facilitator School*. https://www.facilitator.school/blog/what-is-a-workshop

www.recruiter.com. (2023, April 26). *Pros and Cons of Using AI in Recruiting*. Recruiter.Com. https://www.recruiter.com/recruiting/pros-and-cons-of-using-ai-in-recruiting/

YLE. (2023, November 11). *Loimme 13-vuotiaan Ellan – Tiktok tarjosi hänelle sisältöä itsemurhasta ja kalorien laskemisesta*. Yle Uutiset. https://yle.fi/a/74-20059318

Zhong, Z. (2018, October 22). *A Tutorial on Fairness in Machine Learning*. Medium. https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb

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.

2022-1-ES01-KA220-HED-000085257