Digital Skills & Jobs Platform

Artificial Intelligence: A high-stakes game, but at what cost?

Deep-dive

This paper is part of our digital brief series: a set of easy-to-grasp deep-dives, tackling the latest key topics around digital skills and jobs, produced in collaboration with some of Europe's best experts in the field.

2024

The Digital Skills and Jobs Platform is funded by the Connecting Europe Facility of the European Union. Opinions expressed do not necessarily reflect the contracting authority's official position.

Table of Contents

Figures and tables

Summary

This paper provides a high-level introduction to the topic of artificial intelligence (AI), including examples of how AI and the underlying technologies have been used in practice to predict, optimise, and support human actors in their decision-making. Whilst AI systems have the potential to contribute positively towards improving societal and environmental challenges, this paper also shines a light on some of the many threats and ethical challenges that are firmly part of our discussions of AI. Such challenges include the lack of explainable AI, the extensive use of training data and human AI labelling labour, the creation of deepfakes and misinformation, and the negative environmental impact from the fast-paced acceleration of large-scale AI systems. With the longer-term positive and negative implications and consequences of AI yet to appear on the horizon, regulation has entered the arena to support and create guardrails for the design, development and deployment of AI systems. This paper focusses on the efforts by the European Union through the EU AI Act, before considering how societies can enable an upskilling of the workforce towards having the capabilities and competences to work with and alongside AI.

Keywords

Artificial Intelligence, Innovation, Ethical challenges, AI Regulation.

Author's biography

Dr Rikke Duus is an Associate Professor at University College London (UCL) and a member of the visiting faculty team at ETH Zurich. She is part of the UCL Academic Board and works with EFMD to certify digital and AI-enabled programmes. As a member of several of the European Commission's Digital Education Working Groups, she collaborates with education, policy and EdTech teams on AI and digital education and has presented at EU-organised workshops in Brussels and Paris. She is also a contributor to the development of the European Commission's General-Purpose AI Code of Practice. Rikke undertakes research in the areas of digital and AI-enabled transformation, ecosystem-based value creation, and UN SDG-led innovation. She has won several awards for her work in digital education and research. Her work is published in academic journals, practitioner outlets and global media, including the World Economic Forum, Frontiers in Psychology, Harvard Business Publishing Education, and The European Business Review. This year, she has contributed to a report on the Future of Education and Skills to be published by the OECD and written a paper for the European Digital Education Hub on Fluid Hybrid Learning Spheres. Rikke is also the co-architect of the award-winning team-based innovation programme, the Global DigitalHack®. She has over 15 years of executive education experience, working with public and private sector organisations and industry leaders.

Introduction

Although we might not realise it, every day, we interact with services or devices that are powered by artificial intelligence (AI). It can be anything from doing your online shopping and receiving personalised recommendations, interacting with chatbots and being informed of real-time shipping and delay estimates, to receiving live route guidance from apps like Google Maps, Apple Maps and Waze. Services like these use relevant data sources, e.g. past purchase data, browsing data, location and weather data, that AI systems can interpret to help deliver the best possible user experience. Most people might find these services helpful. So why is AI a high-stake game? And what are the costs? As with most technologies, there will always be a debate to determine if the impact of such technologies can contribute to the betterment of society, people and planet, or impact adversely to the detriment of all. This is especially the case with technologies that evolve fast, such as AI, and where the longer-term and deeper-rooted consequences are yet to fully emerge. Whilst AI is not a new phenomenon, the world has experienced an acceleration in the development of different types of AI-enabled systems and interfaces in the last few years.

Before we venture deeper into the world of AI, it is important to firstly acknowledge that the field of AI is fast-moving. Therefore, the information, insight and examples provided are contextualised to the time of writing. Second, AI has become somewhat of a catch-all term for the many and varied technologies that enable the creation of AI systems and hardware (e.g. robots) that attempt to mimic human behaviours, including reasoning, learning, problem-solving and decision-making. The engines behind AI include machine learning, deep learning, natural language processing, computer vision and sensor technology amongst other technologies that can use data to understand, interpret, predict and decide on actions aligned to set objectives. Throughout this article, reference is made to AI systems and AI-enabled systems. It is acknowledged that functioning AI solutions are reliant on the support of effective machine learning processes (including deep learning and natural language processing) that can learn from reliable and relevant data (including text, audio, images, video), to identify patterns and make decisions with some or minimal human intervention. Where it is possible and meaningful, reference has been made to the underlying technology(ies) supporting the AI systems in the examples to follow. Where the terms *'AI system'* or '*AI-enabled system*' have been used, it should be assumed that machine learning and other data analysis methods have been employed to help build decision models and machine learning algorithms to enable the AI applications.

In the next couple of sections, we explore further what AI is and how AI systems can be designed. We then turn to the darker sides that also follow in the slipstream of AI and consider some of the ethical challenges when designing, developing and deploying AI systems.

Decoding Artificial Intelligence

The field of AI is constantly debated, fuelled by an extensive media coverage and the frequent updates from Big Tech about the latest capabilities their AI systems possess. Perhaps like no other technology, it has split the waters and created camps of both "techno-optimists" and "techno-sceptics" (Heaven, 2024); and it would seem there are good reasons for this. Consider these 3 examples: in the education sector, learner data and analytics can be used to create personalised learning pathways to support individual needs; but can also be seen to create a data-driven surveillance culture (Selwyn, 2024). In the news and

media sector, AI copy-writing tools can turn minutes from a council planning committee meeting into a news report in the publisher's style, freeing up journalists to do more investigative work in the field. At the same time, AI has enabled hundreds of fake news sites to pop up around the world almost overnight. Often using AI chatbots to paraphrase articles from other news sites with little to now human oversight, many of these sites are created to deliver *'auto-generated misinformation'* to influence societal debates and citizen beliefs and perspectives on critical issues, e.g. who to vote for in an upcoming election (France 24, 2024; Hill & Hsu, 2024). In some parts of the healthcare system, machine learning algorithms are already used to analyse and evaluate patient test results, undertake needs-based prioritisation of patients, and give input on patient treatment and care plans. The risks of bias derived from the datasets the AI systems are trained on, ensuring equitable access to AI systems that do enable positive patient outcomes, and safekeeping of patients' private data (Solanki, Grundy & Hussain, 2023), are, however, ever-present challenges. All in all, AI is a field marked by great levels of anticipation and hesitation at the same time.

Before the emergence of advanced large language models (LLMs), such as GPT-4, Llama, Gemini and Claude, the role of AI was more focussed on applying decision-algorithms to help make predictions and recommendations based on a user's or other sources of data, or to engage in relatively simple conversations to help, support or provide information to users when prompted – see Figure 1 below (Hermann & Puntoni, 2024).

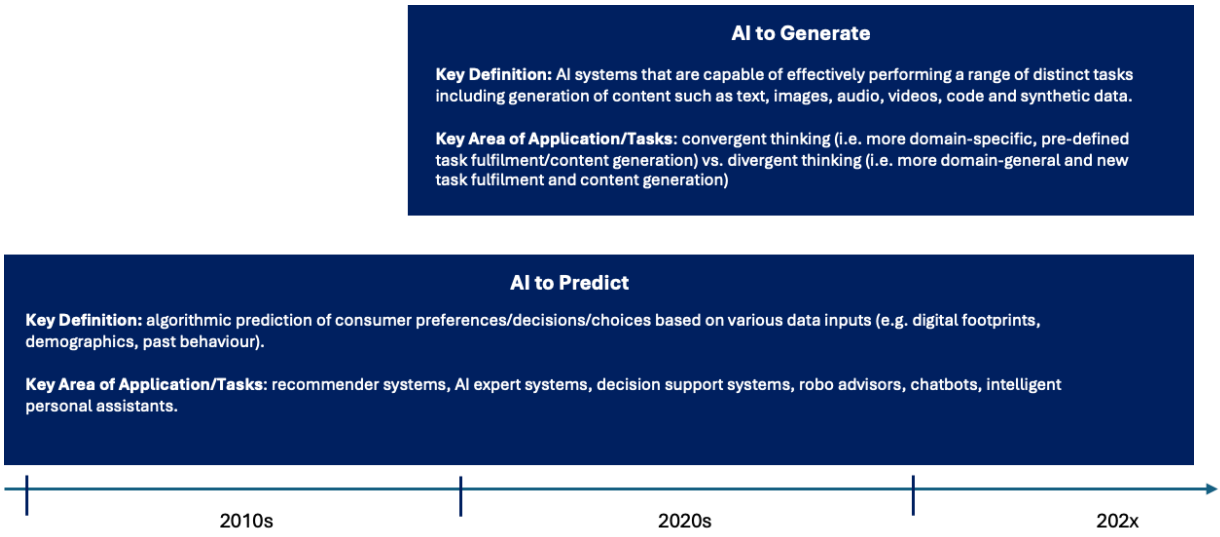
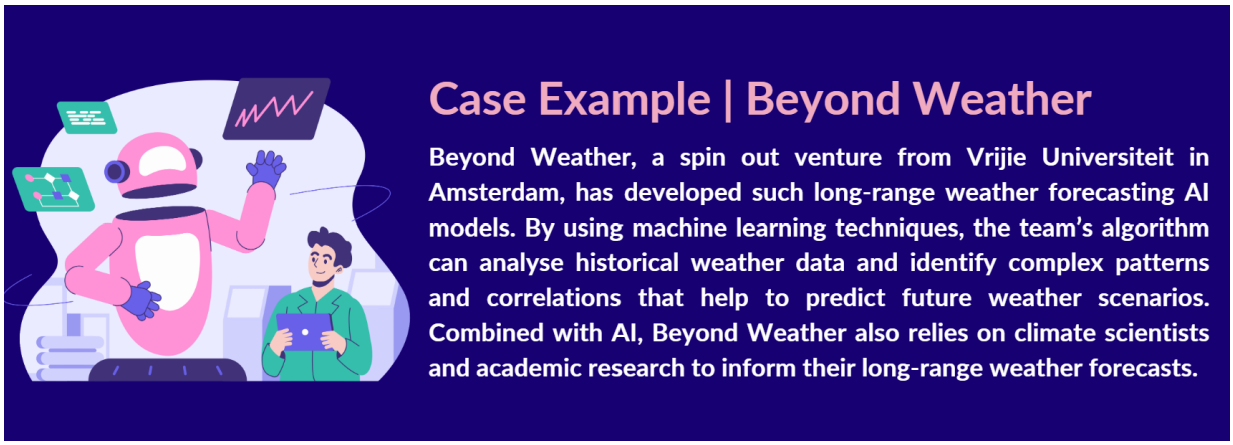


Figure 1. From Predictive AI to Generative AI | Source: *Hermann and Puntoni, 2024 (Please note, this figure has been adapted from the original source, available at: https://doi.org/10.1016/j.jbusres.2024.114720)*

An algorithm can be described as a process that has been automated and given the ability to make decisions independently without requiring input from a human, using data, statistics and other computing resources (Mahmud, Islam, Ahmed & Smolander, 2022). In other words, an algorithm can be defined as "*a set of steps that a computer can follow to perform a task*" (Castelo, Bos & Lehmen, 2019, p. 809). This includes AI models that are designed to create long-range weather forecasts, which can enable organisations in the energy and agricultural sectors to better plan ahead by anticipating high-risk weather events even earlier than what is currently possible. *Beyond Weather* is a start-up pioneering in this area.



Case Example 1. Beyond Weather

Interestingly, some people express a behaviour referred to as "*algorithmic aversion*" (Dietvorst, Simmons & Massey, 2014). This happens when a person chooses, consciously or subconsciously, to reject or disregard the recommendation or advice provided by an algorithm; especially if the algorithm has made a mistake in the past (e.g. given a poor recommendation or the wrong directions). This might represent only a minor issue when following a GPS's location directions - but in healthcare, for example - it is critical that medical staff can trust an AI system's guidance and recommendations.

Building on AI's ability to predict, we have now entered the phase where some AI systems possess generative abilities (Hermann and Puntoni, 2024, Figure 1). Generative AI has the ability to produce content (i.e. images, text, sound, video, code and data) and operate with an increased level of autonomy. LLMs are typically the engines under the hood of these AI systems. By design, LLMs are based on deep learning algorithms that use very large data sets to undertake pattern recognition, translation, predictions, summarisations, and generation of new content. AI systems that draw on sensor-based data are already used to create digital twins of cities, such as Helsinki, Barcelona and Singapore, and the next phase is to use generative AI systems.

Digital twins & CitiVERSE

Digital twins are real-time digital models of the real world, e.g. an urban environment. A digital twin of a city requires access to relevant data about movements and activities within the city, typically through sensor technology, as well as historical records and environmental conditions. This data is then input into the AI models and machine learning algorithms that make sense of the vast data to identify patterns, detect anomalies, and predict challenges. Digital twins of urban environments can encourage transparency and accountability amongst key stakeholders, including citizens, when making decisions about the urban space.

The European Commission has launched the CitiVERSE initiative, which aims to connect existing local digital twins across Europe to form the EU CitiVERE. From a technology perspective, the initiative is focused on advancing generative AI applications, including enhanced simulations to address air quality, decarbonisation, congestion, and improve citizen interaction amongst other critical societal and environmental challenges.

Some of the key characteristics of generative AI includes its ability to adapt and evolve, perform in areas it was not intentionally trained for, being able to learn from relatively limited data and being able to learn over time (Triguero et al., 2023). One of the central differences between systems that do not rely on AI and those that do, is that systems without AI require a human to take the lead to set the factors needed for the algorithm to make decisions, whilst in AI systems, algorithms have greater autonomy to make inferences based on pattern recognition and generative abilities (Mahmud, Islam, Ahmed & Smolander, 2022).

To guide and influence the evolution of AI, more than 20 countries in Europe, including Germany, Denmark, Spain and Slovakia, have developed their own AI strategies to encourage positive innovation and equitable societal impact from AI, whilst also putting in place guardrails and protective guidelines to limit the negative impact. From the perspective of the European Union, two challenges that need careful consideration is how to ensure that AI solutions are developed and deployed in an ethical and equitable manner, and how to ensure the education of qualified professionals who possess relevant skills, competences and expertise to design, develop, deploy and manage new AI solutions. Both of these challenges are explored later in the article, but for now, we turn our attention to how organisations, public sector bodies, tech companies and startups can go about designing AI systems.

Designing AI Systems

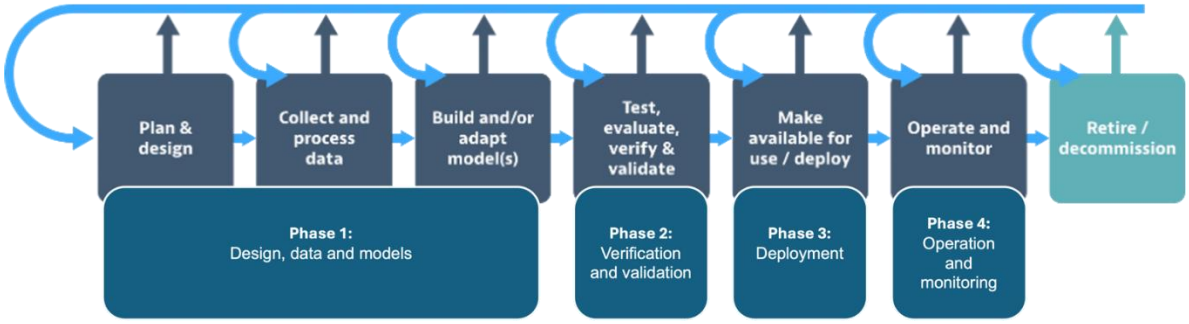
The OECD presents the AI system lifecycle which provides a pictorial illustration of the key phases that often take place in the design, development and deployment process of AI systems (Figure 2). The four phases include:

- Phase 1 | Design, data and models
- Phase 2 | Verification and validation
- Phase 3 | Deployment

► Phase 4 | Operation and monitoring

The approach to the design and development of an AI system will depend on the specific context that the AI system is being built for and the specific context that it will be deployed within. This will also impact what data the AI system is given, how the data is processed and analysed as well as the specific algorithmic mechanisms of the AI model(s). Whilst the illustration in Figure 2 may indicate that this is a linear and step-by-step process, planning, designing, building and deploying AI systems will often be an iterative process – and the decision to retire or decommission an AI system can happen across the four phases of the AI system life cycle. In July 2024, McDonald's brought to an end its 3-year collaboration with IBM to test an AI-based automated order system at over 100 drive-thru McDonald's restaurants across the US. No official reason was given for the shut down.

Figure 2. AI system lifecycle | Source: OECD AI Principles Overview

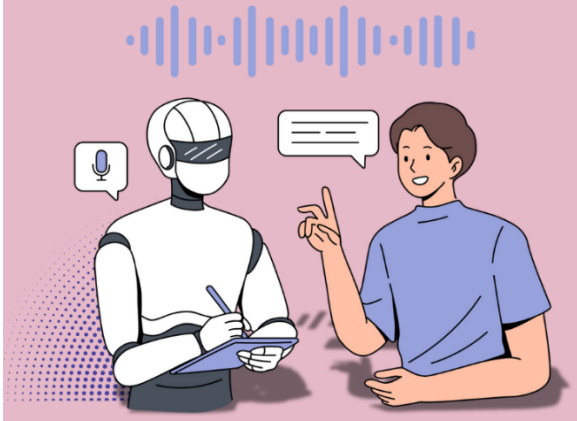


The level of autonomy that is built into an AI system will depend on the specific scenario within which the system is intended to operate. Here, it is helpful to think about three levels of AI systems in terms of autonomy and human involvement, namely, 'Human-in-the-loop', 'Human-on-the-loop' and 'Human-out-of-the-loop'.

'Human-in-the-loop'

In these AI systems, a human is meaningfully involved in overseeing and instigating the actions of the AI system (EU-U.S. Terminology and Taxonomy for Artificial Intelligence, 2024). For example, in a healthcare setting it may be desirable and preferrable to have a human clinician as the main decision maker supported by input from the AI system.

EXAMPLE



Case Example | Primage

This is the approach adopted by the team behind the EU-funded PRIMAGE project that was created to assist diagnosis, prognosis and therapy in children with aggressive brain tumours. The research team utilised different types of retrospective data, e.g. imaging, clinical, molecular, and genetics, from paediatric oncology units across Europe and the European Society for Paediatric Oncology to design a decision support tool based on AI models. The main objectives were to develop the capability to predict clinical outcomes and support clinicians to make personalised decisions for their patients (Primage, 2024).

Case Example 2 - Primage

'Human-on-the-loop'

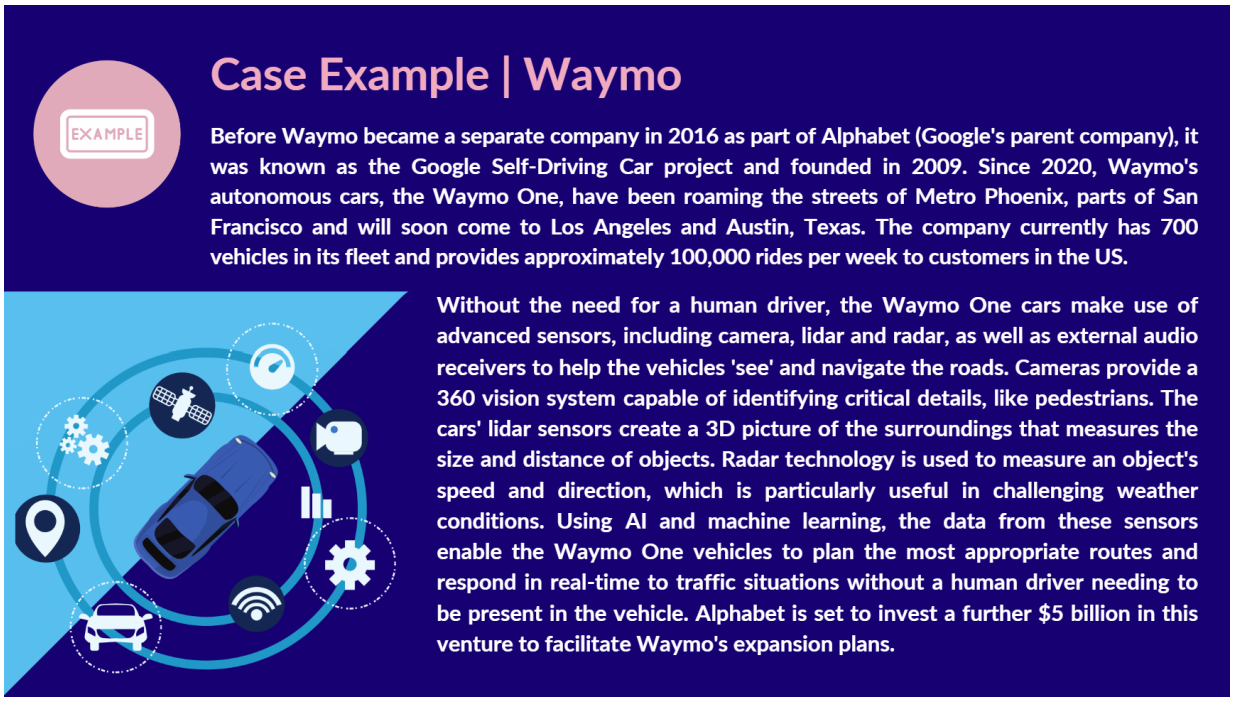
Here, a human is present to check the actions and decisions by the AI system and with agency to also abort intended actions the AI system has put forward. This approach is typically used when the AI system has obtained a certain level of performance, albeit still needs some human interaction and input.

CASE STUDY

An example of 'Human-on the-loop is DHL's last mile delivery route optimisation. Using forecasting and prediction models, DHL has 90-95% clarity on where and when shipments will arrive at specific facilities. With this insight, the routes for couriers are planned, considering volumes, service and other influencing variables. The delivery route plans that have been put together are then further optimised via AI-enabled software, which is capable of creating the most optimal sequence of stops taking into account the level of urgency of delivery (e.g. an urgent medical express delivery) and distance per stop. Customers receive a predicted time of delivery, which is updated as the courier gets closer to the customer.

'Human-out-of-the-loop'

For fully autonomous AI systems where no human action is involved, these systems can be referred to as '*human-out-of-the-loop*'. AI systems with a high level of autonomy have the ability to learn, adapt, analyse and respond to emergent situations that have not been anticipated and therefore have not been pre-programmed within the system.



Case Example 3. Waymo

Data is essential for the development of AI models and systems. In fact, AI models are "extremely dependent on data collection, which in the context of our current internet landscape has contributed to the rise of 'surveillance capitalism'" (Krasodomski, 2024, p. 50). Data is needed to train machine learning models to enable them to learn from information and develop generative capabilities. Data is also used to assess and test a model's reliability, accuracy and validity, which often occur in Phase 2 (Verification and validation) of the AI system lifecycle. There are different types of data that can be used to train machine learning models and AI systems (Snaith, 2023), such as:

- Textual data: This includes text-based information. CommonCrawl's extensive archive of textual data has been used in several training models, including OpenAI's GPT-3.
- Visual data: This includes image-based data. The first Stable Diffusion model was trained on more than 2.3 billion image-text pairs spanning a wide range of topics.
- Synthetic data: This is data that is artificially created, typically by using algorithms that draw on real-world data. Synthetic data may be used when real-world data is not accessible, where it is of a sensitive nature or where the available data is not adequate to develop a reliable AI-model.

As explored in the section below, one of the significant ethical challenges relate to how data is being sourced and used for training of AI models and the impact of these practices on an AI system's agency, decision-making responses and presentation of information.

Ethical Challenges of AI

There are many examples of how AI systems and the supporting technologies can be used for good to enhance societal value and wellbeing. However, there are also many unignorable and visible threats that make AI development and adoption a high-stakes game. For example, researchers at MIT have identified more than 700 risks from AI systems across seven core areas (MIT, 2024). Many of these threats and ethical challenges have already become visible in our societies today. For the purpose of this article, we focus on threats and ethical challenges that broadly fall into the three domains in *Table 1. 'Technical', 'Societal' and 'Environmental' Domains*

- Technical – the design, development, training and testing of AI models and systems
- Societal – the impact of AI models and systems on individuals, citizen sub-groups and society at large
- Environmental – the impact of AI models, systems and hardware on the environment

Table 1. 'Technical', 'Societal' and 'Environmental' Domains

DOMAIN	THREATS AND ETHICAL CHALLENGES
Technical	Threats and ethical challenges in the Technical domain are primarily derived from: ▶ How the AI models, algorithms and AI applications are designed and applied ▶ How and from where the training data has been sourced ▶ How the AI models have been trained to avoid bias and hallucinations ▶ The extent to which the AI systems are designed to be transparent with contextualised explainability ▶ The lack of transparency of AI-generated content Some threats and ethical challenges transcend more than one domain. For example, AI labelling labour is both a Technical and Societal concern. AI labelling is currently part of the development process of many AI systems and LLMs, whilst also causing a concern over the exploitation of AI labelling workers in the Global South.
Societal	Threats and ethical challenges in the Societal domain are primarily derived from the broader impact of AI on citizens, citizen groups, and individuals. These threats and ethical challenges are derived from: ▶ An increase in AI-generated content, such as deepfakes and misinformation ▶ Exploitation of organisations' security vulnerabilities, leading to hacking and cyber attacks ▶ A reduction of data privacy as Big Tech and other AI leaders extract data to train their AI ▶ Erosion of trust from algorithmic bias ▶ Discrimination against certain citizen sub-groups Some threats and ethical challenges transcend more than one domain. For example, inequality is both a Societal and an Environmental concern. In some regions of the world, data centers can be operated with a greater proportion of

	carbon-free energy, while in other parts of the world, energy comes from fossil fuels. This disproportionality adversely impacts the level of air pollution and carbon emissions generated from data centers and consequently affects human health and wellbeing.
Environmental	Threats and ethical challenges in the Environmental domain are focused on the environmental costs of the accelerated design, development, deployment, and maintenance of AI systems. The negative impact on the environment is primarily derived from: ► The vast amount of large-scale data centres needed to store and process training data for machine learning models and LLMs ► Data centres consume extensive amounts of electricity, whilst water is used to cool down computing equipment and in most forms of fuel and power generation ► Inputting a query in an LLM, such as ChatGPT, which requires more electricity than to process a search in a search engine, like Google, Bing or Mozilla Firefox ► The manufacturing of AI chips and AI processing hardware equipment Some threats and ethical challenges transcend more than one domain. Data centers and chip manufacturing are both Technical and Environmental concerns. For example, chips manufacturing has a negative impact on greenhouse gas emissions, energy and water consumption, whilst raw materials are used for manufacturing processes (e.g. palladium, copper, cobalt and rare earth materials).

It is important to highlight that there are overlaps, interactions and transfers of impact across the three domains. In fact, many of the threats and ethical challenges that arise in the Societal and Environmental domains can be traced back to how the AI models and systems have been designed, trained and tested, i.e. activities that belong to the Technical domain. With the presence of the three domains, Figure 3 provides a high-level illustration of some of the identifiable threats and ethical challenges assigned to a primary domain. The threats and ethical challenges identified in Figure 3 is not a definitive or exhaustive list, but provides us with a helpful illustration to pinpoint some of the impact areas that need attention as big tech, organisations, governments and other actors speed ahead in their development of more and more advanced AI systems.

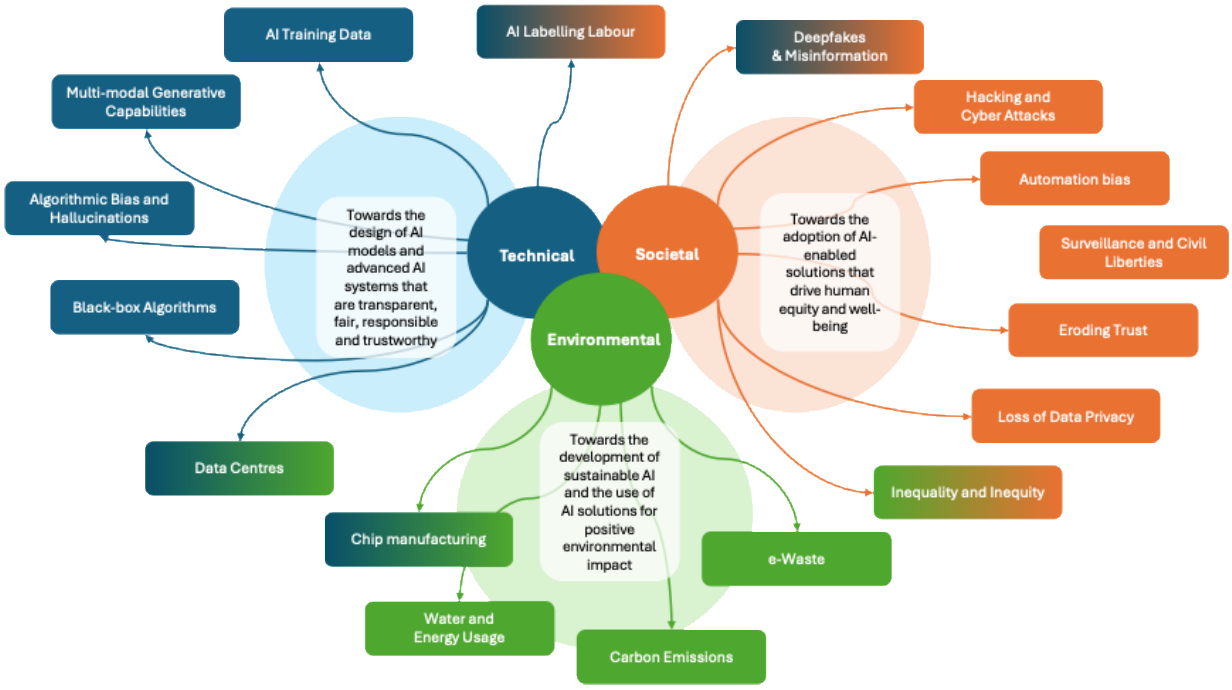


Figure 3. AI Threats and Ethical Challenges | Source: Author

In the following sections, the threats and ethical implications related to blackbox AI, AI training data, AI labelling labour, deepfakes and misinformation, and AI's environmental impact are explored in some further depth.

Overwhelmed? Take a look at the infographic below, which illustrates some of the main concepts we introduced today.

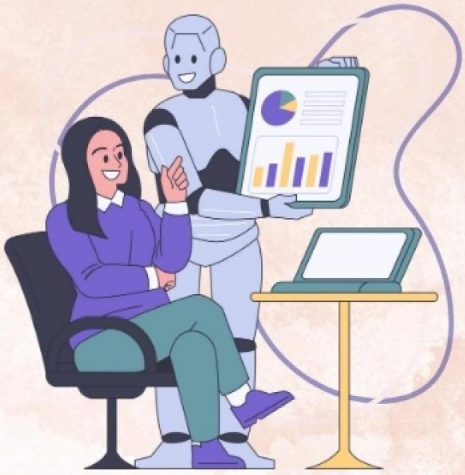
Artificial Intelligence (AI): a high-stakes game - but at what cost?

AI SYSTEMS

AI has become a catch-all term for many technologies that attempt to mimic human behaviour, including reasoning, learning, problem-solving and decision-making.

The engines behind AI include machine learning, deep learning, natural language processing, computer vision and sensor technology amongst other technologies that can use data to understand, interpret, predict and decide on actions aligned to set objectives.

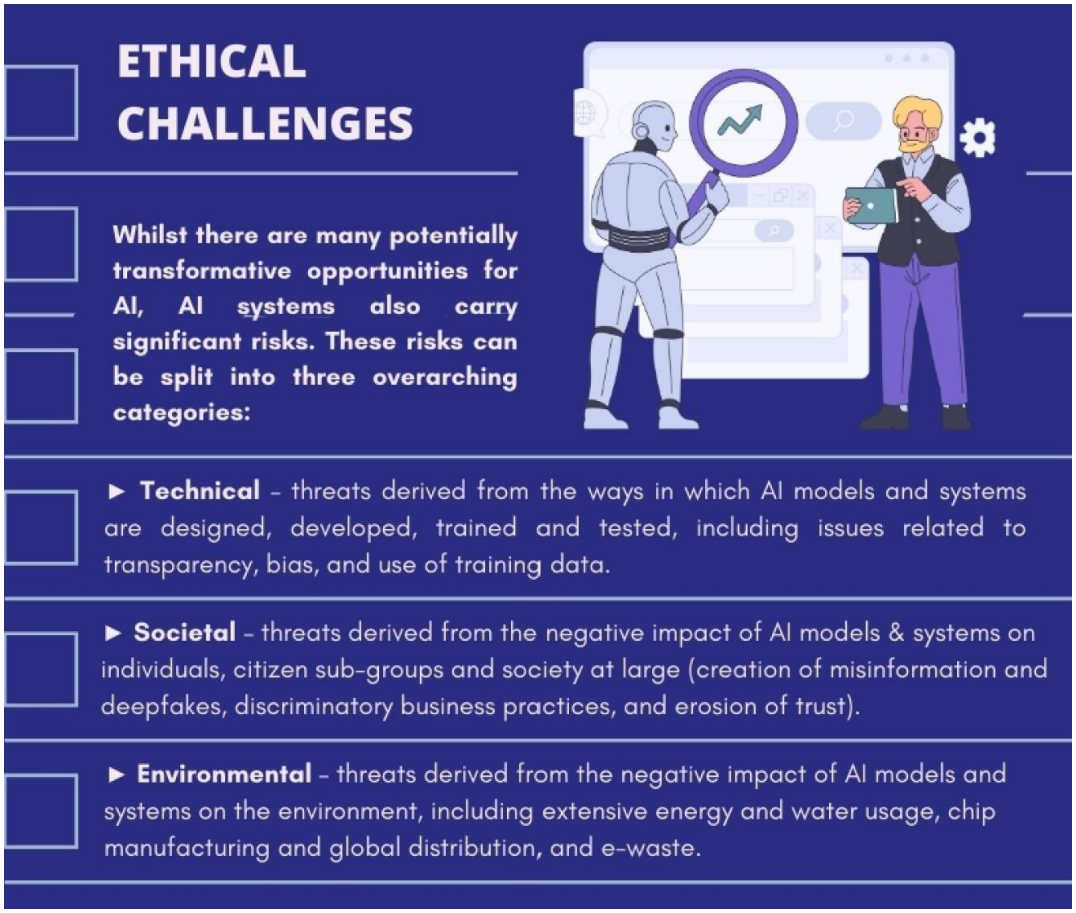
To function, an AI solution is reliant on the support of effective machine learning processes (deep learning, natural language processing) that can learn from data (text, audio, images, video) to identify patterns and make decisions with some, or minimal, human intervention.



ALGORITHMS

Algorithms are the core mechanisms that enable AI systems to function. In AI, algorithms are the instructions that guide machines on how to behave intelligently. An algorithm can be described as a process that has been automated with the ability to make decisions independently without requiring input from a human, using data, statistics and other computing resources - essentially, a set of steps a computer can follow to perform a task.

This includes AI models that are designed to create long-range weather forecasts, which can enable organisations in the energy and agricultural sectors to better plan ahead by anticipating high-risk weather events even earlier than what is currently possible.



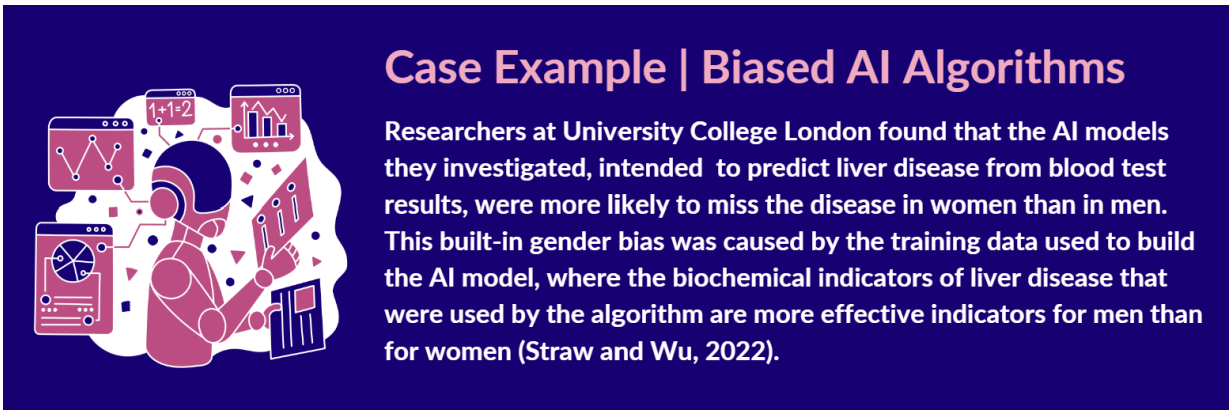
Blackbox AI

Much of the time, AI systems suffer from the '*blackbox issue*' (Davenport & Ronanki, 2018). This means that it is not obvious to the receiver of the AI system's recommendations how these were arrived at due to a lack of explainability and transparency of the variables used for consideration, the data sources and input, and the decision criteria. This can lead to doubt and mistrust in the recommendations put forward (Borges et al., 2021) and a feeling of unease about following the AI system's proposed actions. Some people even report experiencing 'technostress' when they do not understand or feel knowledgeable about the reasoning processes that an AI system uses to make decisions (Issa et al., 2024). As AI systems become more complex and seek to draw on a greater variety of historical and real-time data sources, heightening the transparency and explainability of AI becomes ever more critical. AI systems that are introduced with the intention of employees adopting them to enhance their decision-making, creativity and efficiency, should have a good level of awareness of what data the system draws on, how the data is analysed, the variables applied and the accuracy with which the AI system makes its outputs. If not, there is a risk that people become *'slaved to the system'*, as they are unable to decode and understand how the AI system operates, how it develops it recommendations and makes its decisions. This is especially important in fields such as healthcare, recruitment, and policing, where peoples' lives can become adversely impacted, especially if decisions made prove to be wrong or inaccurate. In September 2024, Europol published its report on AI and policing, exploring both the benefits and challenges of using AI in law enforcement. Some of the challenges highlighted include data bias and fairness, privacy and surveillance, and accountability and transparency. AI systems used by law enforcement fall for profiling or risk assessments fall within the

high-risk category as set out in the European Union's AI Act that came into force on August 1st 2024 (see section on European AI Regulation). High-risk AI system need to be transparent and explainable so that they can be trusted to behave in reliable ways and as intended.

AI Training Data

All AI models need access to relevant training data to develop their predictive and/or generative competences. Thinking about AI models with predictive capabilities, the expectation is that these models will be capable of analysing and evaluating data towards making recommended decisions and assisting human decision-makers. These AI models are often built for a specific purpose; for example, AI models that are built to assist medical teams with analysing large numbers of x-rays of patients who may have breast cancer to efficiently detect those who display symptoms and need care. In these use cases, it is especially important that the AI model's recommended decisions are accurate, reliable, consistent and fair. However, in developing AI models for medical care there are examples of systems that underdiagnose a particular subset of the population.



Case Example 4. Biased AI Algorithms

Essentially, if an algorithm is trained on poor, insufficient or inappropriate data, its outputs will also be low in quality and may lead to unreliable, biased and inaccurate outputs that could end up significantly disadvantaging some parts of the population. For companies like Google, Meta and OpenAI who have all launched LLMs, the focus is on getting access to as much data as possible to continue to develop their LLMs. Already, huge amounts of data available on the internet have been used to train their LLMs, though not without facing legal challenges over the use of copyrighted material (Metz, Kang, Frenkel, Thompson & Grant, 2024).

Case Example | Copyright infringement

In December 2023, the New York Times launched a lawsuit against OpenAI and Microsoft for copyright infringement over the use of millions of articles, published by the news media, to train AI technologies, including ChatGPT. This is just one of several lawsuits filed on the grounds of copyright infringement, including lawsuits by renowned writers John Grisham and George R.R. Martin (Games of Thrones), as well as stock-photo provider Getty Images.

Case Example 5. Copyright infringements

The largest language models are trained on datasets with tens of trillions of words and these datasets double approximately every eight months (tokens = words or pieces of words), as illustrated in Figure 4.

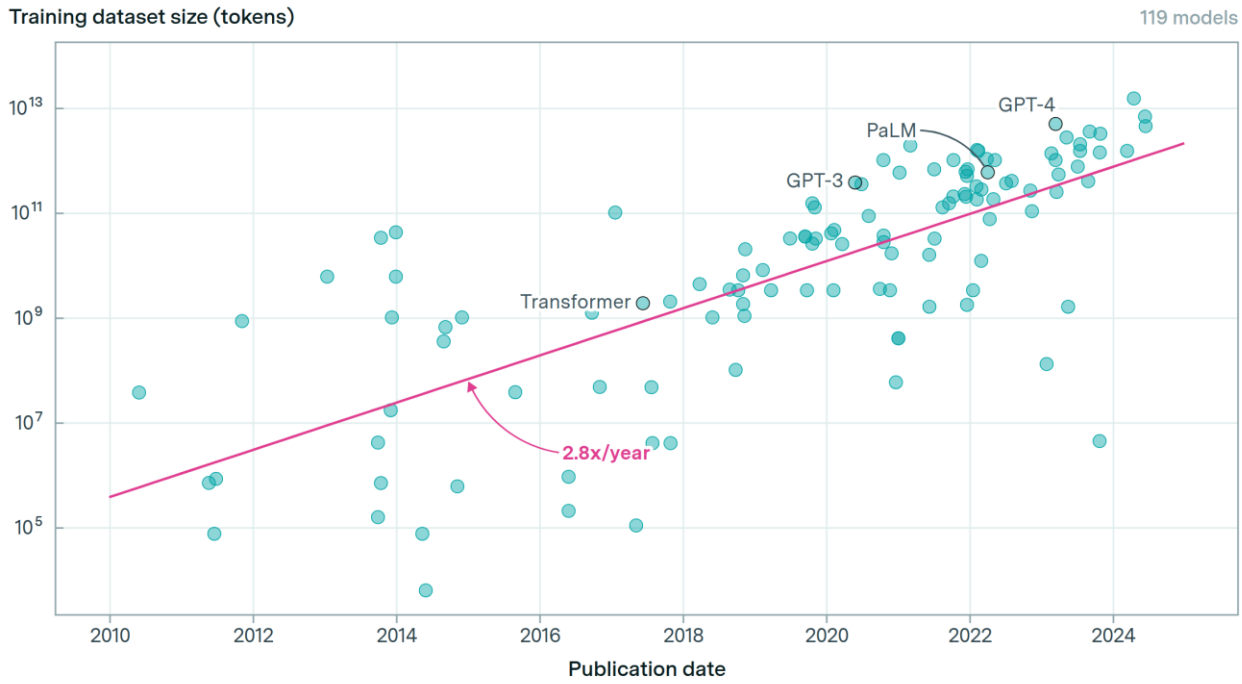


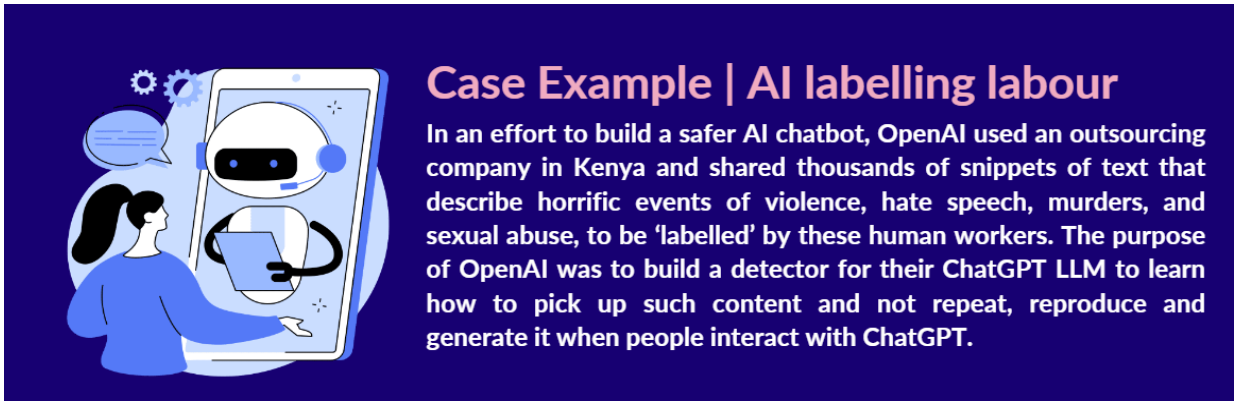
Figure 4. Training data of notable LLMs | Source: Epoch AI, 2024.

Problematically, tech companies are using human-created data to train their AI systems faster than it is being produced, which poses an ethical risk now and in the future. In fact, public human text data may run out already at some point between 2026 and 2032 (Villalobos, Ho, Sevilla, Besiroglu, Heim, & Hobbhahn, 2024). There are opportunities, however, for AI developers to overcome this challenge by improving algorithms to require less training data, use a greater amount of synthetic data, or train the algorithms to learn from new types of data (Matulionyte, 2023). As we might expect, these options are not without their own ethical challenges. For example, if future AI models are to rely on synthetic data, i.e. data created by AI, it could lead to an exacerbation of racial, gender, age and other forms of bias, which are already present in much of the human-created '*dirty data*' that today's AI models are being trained on. Added to this ethical issue, generative AI also hallucinates and makes up false and inaccurate information, which is fed back in to the pool of data, creating a potentially harmful feedback loop.

AI Labelling Labour

The human work that is involved in building and maintaining AI models and algorithms is not typically visible or part of Big Tech's conversations about accelerating innovation in AI. Yet, human labour is a significant cog in the wheel that enables tech companies to build ever larger and more advanced AI systems and LLMs. Two of the roles that human workers take on behind the scenes are as '*data annotator*' and '*content moderator*'. The primary role of data annotators is to label data with 'tags', which enables computer programmes and machine learning models to understand and '*read*' the information (Muldoon, Cant, Wu & Graham, 2024). Content moderators are tasked with viewing and reviewing content and removing any content that is deemed to be harmful and in breach of company guidelines. The role of content moderators and the often traumatic psychological impact of this job has already been brought to light in the mainstream media and academic research in the context of social media platforms, such as TikTok and Facebook (see for example, Pinchevski, 2023).

Returning to the role of data annotators who label and tag data, this workforce is typically engaged during the pre-training process of an AI model. During this phase, the AI model is taught how to generate outputs based on data sets that have been labelled by these human annotators. The two main objectives are to improve accuracy of the AI model and to enhance safety, which could include reducing the risk of infringing on copyrights, minimising the amount of misinformation and overcoming bias (Taylor, 2023). Whilst these are positive objectives to strive for, AI labelling labour also reveals the underbelly of AI systems that are marketed and positioned to the public as 'automated' and in essence 'human free' (Bartholomew, 2023). In reality, however, cheap human labour is sought from, amongst other, people in the Global South to get the data labelling work done.



Case Example 6. AI labelling labour | Source: Perrigo, 2023.

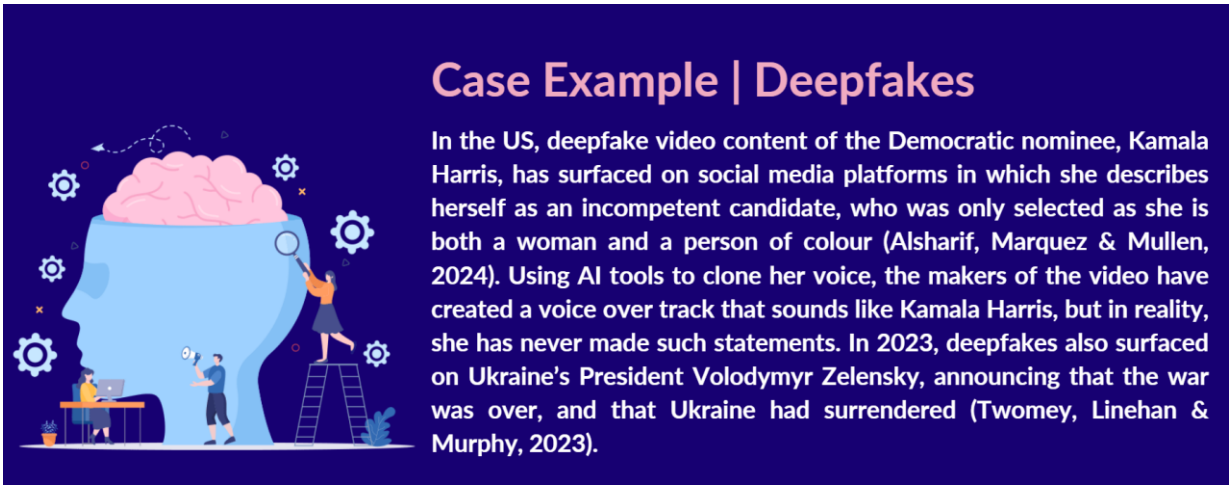
Whilst training the LLM to know that this kind of content is 'bad', there is a significant human cost for the many people who become exposed to it as part of their labelling work and without much financial gain or reward (Muldoon, Graham & Cant, 2024).

Deepfakes and Misinformation

A deepfake is a piece of content, e.g. an image, audio or video, that has been digitally manipulated using AI tools to make a person or persons say or do something that never actually happened. A deepfake can also be of an object or event, such as the AI-generated image of an explosion next to the Pentagon in the US in 2023 that never actually took place in real life (Hurst, 2023). Deepfakes become particularly dangerous when they are designed to create harm, deceive or mislead as part of misinformation efforts.

With the quick advancement in AI tools, deepfake videos can be created from a snippet of a person's voice or even from just a still photo of the person's face. Social media platforms can then be used for quick distribution of the content to reach millions of viewers.

There are positive uses of this technology too. For example, when it is used to enable people who have lost the ability to speak due a medical condition or treatment to hear their own voice again. However, currently, the fear is perhaps greater than the positive benefits as deepfakes are already influencing and shaping public opinion, perception and action through misinformation. Deepfakes become particularly convincing when they mix real and fake content.



Case Example 7. Deepfakes

As deepfakes become more and more convincing, distinguishing between authentic content and that created and generated by AI, will become increasingly challenging, which can further deteriorate the public's trust in information. This is supported by a recent study by The Alan Turing Institute (Sippy, Enock, Bright & Margetts, 2024), which found that 87.4% of people in the United Kingdom are concerned about deepfakes affecting election results, and 91.8% are concerned about deepfakes becoming more common place and impact areas such as child sexual abuse material, increasing distrust in information and the manipulation of public opinion.

As the challenge of deepfake content continues to rise, so does the number of companies offering deepfake detection software. Some companies use machine learning techniques to detect the 'fingerprints' (i.e. unique elements also referred to as 'artifacts') in the deepfake content that reveal the use of deepfake AI tools. Despite these efforts, there is currently no silver bullet solution to separate fake from real content as the technology is still in the early stages (Schaul, Verma & Zakrzewski, 2024).

Earlier this year, the European Commission, under the Digital Services Act, requested information from big tech, including Microsoft (Bing), Google (Search and YouTube), Facebook (including Instagram), Snapchat, and TikTok on their efforts to mitigate the use of deepfakes to mislead and deceive. Since then, Google has announced that it will make explicit AI-generated content more difficult to find via its search engine by changing the way search results are ranked. For example, if someone searches on a deepfake nude of a specific celebrity, Google's algorithms will instead look to deliver articles about the threats of deepfake content rather than the explicit content itself (Dave, 2024). AI-generated deepfake content now fall into the *'Specific transparency risk'* category of the EU AI Act, which means that this kind of AI generated content must be clearly labelled as having been generated by an AI tool or system.

AI's Environmental Impact

The environmental impact of AI is an area that has gained more attention in the wake of the continuous growth and expansion of AI systems and LLMs, like ChatGPT, Midjourney and Gemini, which require extensive computing power. The negative environmental impact of AI can be considered in terms of:

- **Computer hardware:** Energy required to manufacture General Processing Units (GPU) chips, which involves intensive mining and disposal (often leading to e-waste). Most LLMs today use GPU chips from one company, namely Nvidia.
- **Training of AI systems:** The actual training of the AI models, and especially generative LLMs and multi-purpose AI systems, also requires extensive computing power and data centres, which leads to a very high energy and water usage, and carbon emissions (Luccioni, Jernite & Strubell, 2024).

Google alone saw its greenhouse gas emissions increase by 48% in 2023 compared with 2019; an increase the tech company attributes to the ever-growing need for energy by its data centres which enable the company's AI efforts (Rahman-Jones, 2024). Data centres also need cooling down, requiring huge amounts of water. This has seen some of the big tech companies in search for locations to build new data centres in more affordable locations and with better electricity rates (Young, 2024). One such place is the state of Iowa in the US, where more than 55% of utility generation comes from renewable energy (54.1% from wind and 1.3% from solar) (Maguire, 2024). Using renewable energy may help tech companies to lower their carbon footprint. Unfortunately, all is not well. In places like Iowa, water is becoming a scarce resource. For 204 weeks, from June 2020 to May 2024, the state experienced areas of moderate to exceptional drought (Iowa Environmental Council, 2024). With Google and Microsoft already having invested in data centres in Iowa, and Meta planning to do the same, this is going to put additional and continued strain on the state's ability to provide access to water for its 3.2 million residents (Young, 2024). Renowned economist Mariana Mazzucato (2024) argues that tech companies should be open and transparent about the resources required before they expand their operation to facilitate their AI acceleration, including the development of new, large data centres.

It is not only the makers of AI models and LLMs that are accelerating the negative environmental impact of this technology; users are also implicated. For example, when requesting a powerful AI model to generate an image, this requires a similar amount of energy to charging a smart phone (Luccioni, Jernite & Strubell, 2024). The wider discussion about how AI systems can be used to help solve the world's environmental grand challenges is therefore somewhat problematic unless AI technology, including machine learning, deep learning, natural language processing and other data-driven tools, become more sustainable, reduce their consumption of essential resources, and identify new, potentially, renewable sources to facilitate their operation.

European AI Regulation

Considering the many threats and ethical challenges derived from AI, it is critically important that these are, not only, recognised, but that ethical and legal frameworks are put in place to help influence the evolution of AI to meet high ethical, transparency and fairness standards. This is especially important considering the fast evolution of generative AI, which presents specific challenges related to exacerbation of bias and prejudices; essentially making AI systems unreliable, untrustworthy, unfair and unequitable

(Krasodomski, 2024). There is not a unified global approach to regulating AI. Some governments are actively working on their own AI regulation or guidelines, with some already in effect, whilst others seek to apply and adapt existing laws to cover the area of AI systems. While regulation is there to safeguard the public, there is also an argument that countries and regions with stricter regulation run the risk of stifling innovation, will become less attractive to existing businesses and start-up ventures in the AI space, and will miss out on the launch of new AI solutions (Espinoza, 2024).

The European Union has introduced the EU AI Act. Approved by the Council of the European Union in May 2024, the law officially came into force on 1st August 2024. This is one of many initiatives by the European Union to guide how the field of AI evolves within Europe and beyond. The EU AI Act provides a global standard for the design, development, deployment and use of AI and applies to areas within EU law, whilst AI systems used for military, defence and research purposes are excluded. The EU AI Act adopts a 'risk-based' approach which means that AI systems are categorised according to the estimated risk associated with their use as presented in Table 2 below (European Commission, 2024a).

Table 2. The EU AI Act's Risk-based Approach

Minimal risk AI systems	Companies face no regulatory obligations under the AI Act, but can voluntarily adopt additional codes of conduct.
Specific transparency risk/ Limited risk AI systems	Companies need to meet specific transparency obligations (e.g. clarity on when users are interacting with a machine and labelling of AI-generated content).
High risk AI systems	Companies are regulated in the development of AI systems that are deemed high risk, requiring risk-mitigation, human oversight, high quality data sets, etc. When an AI system is categorised as 'high risk', the assessment is based on the functions that the AI system performs and the intended purpose and modalities for which the AI system is used. Examples include AI systems that are meant to be used in the management and operation of critical infrastructures, in the determination of access and admission to educational institutions, in the recruitment or selection of candidates for job roles, and in the evaluation of the creditworthiness of individual citizens.
Unacceptable risk AI systems	Companies are prohibited from developing AI systems that are deemed to lead to unacceptable risks to individuals and society. This includes AI systems that are capable of compiling facial recognition databases, exploiting vulnerabilities, evaluating or classifying individuals or groups based on social behaviour or personal traits, inferring emotions in workplaces or educational institutions.

With the EU AI Act in place, Member States have until 2nd August 2025 to create the necessary governance infrastructure to oversee how the rules within the Act are being applied in the development of current and new AI systems. Whilst the rules for AI systems that are deemed to present unacceptable risk will apply after six months, companies have until 2nd August 2026 to ensure they are compliant with the new regulation. To support AI developers with this transition, the EU Commission has set-up the AI

Pact, an initiative that encourages developers to design AI systems that meet the AI Act obligations ahead of the legal deadlines (EU Commission, 2024b). The European AI Office has also launched a multi-stakeholder consultation on trustworthy general-purpose AI models under the AI Act and has brought together AI model providers, industry organisations, civil society organisations, academia and independent experts to participate in the drawing-up of the AI Code of Practice. The AI Code of Practice will be launched in April 2025 and facilitate the appropriate application of the rules of the AI Act.

Upskilling the Workforce for AI

According to the International Monetary Fund (IMF), almost 40% of jobs are likely to be impacted by AI (Cazzaniga, Jaumotte, Li, Melina, Panton, Pizzinelli, Rockall & Tavares, 2024). This number may increase to 60% in advanced economies as many jobs are cognitive-task-oriented jobs. The IMF also estimates that approximately half of those affected may experience a negative impact, whilst the other half are likely to see benefits from enhanced productivity through AI adoption and integration.

Stephany and Teutloff (2024) explain that with technological change, citizens need to review their skillsets as different technologies will require different and novel skills and competences. The research found that AI skills, including deep learning and knowledge of python, have increased in value over recent years. However, it is challenging to know which exact skills are needed for current and future jobs affected by AI. This represents a high level of uncertainty for employers, employees and educational institutions in terms of identifying the skills needed and subsequently creating effective re-skilling programmes. Some companies are now offering AI training for their new recruits. Recently, the American investment bank, JPMorgan Chase announced that all new members of staff would receive '*prompt engineering training*' to more effectively interact with AI systems and applications by writing effective text prompts to generate relevant outputs. The purpose is to boost employee productivity, increase enjoyment of work and add to the company's revenues (Murphy, 2024). Global furniture company, IKEA, has started training approx. 30,000 of its employees and 500 leaders in AI literacy through tailored courses, such as AI Fundamentals, Responsible AI, Mastering Gen AI, and Algorithmic Training for Ethics.

In 2024, the Artificial Intelligence Skills Alliance (ARISA, 2023) project consortium published the AI Skills Strategy for Europe. The strategy contains seven main strategic objectives that set out specific activities, milestones and KPIs. Figure 5 below summarises these seven objectives:

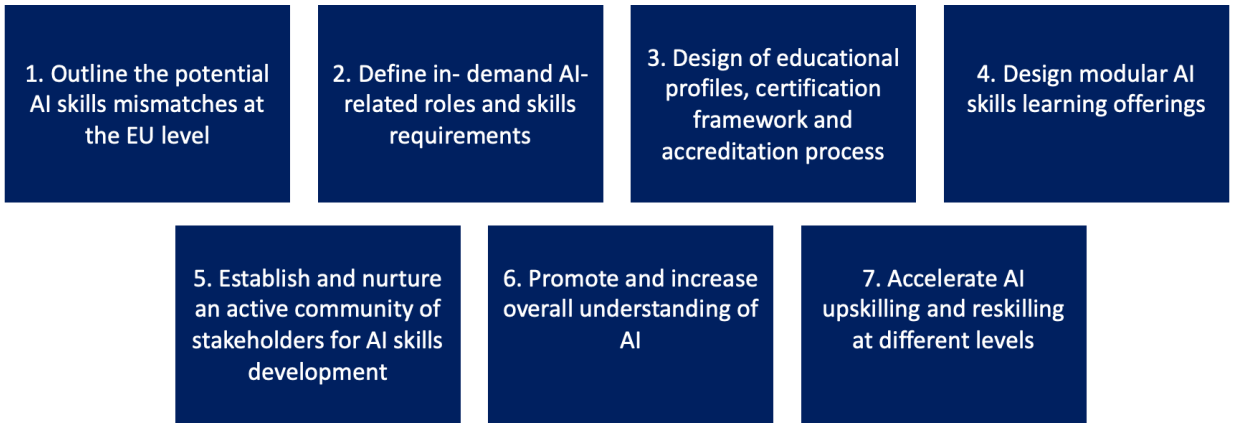


Figure 5. AI Skills Strategy Objectives

A greater focus on how to upskill, at least, part of the workforce to be able to work with and use AI technologies is based on the anticipation that more and more organisations will integrate AI in some way or the other and therefore will need a workforce that is capable and confident to design, deploy and manage AI solutions. Data from the Digital Economy and Society Index (2023) relating to European citizens' digital skills levels show that 26.46% of Europeans believe they have acquired *'above basic'* digital skills in each of the following five dimensions: information, and data literacy, communication and collaboration, problem solving, digital content creation and safety. Whilst having above basic digital skills is a good foundation to build on, specific training may need to be developed to equip citizens and employees with the ability to work with AI. This may indeed help to balance the stakes at play.

References

Arisa (2023). AI Skills Strategy for Europe. *Arisa*. https://aiskills.eu/resource/ai-skills-strategy-for-europe-2/

Bartholomew, J. & Mehta, D. (2023, May 26). How the media is covering ChatGPT. C*olumbia Journalism Review*.

https://www.cjr.org/tow_center/media-coverage-chatgpt.php#:~:text=OpenAI%20launched%20ChatGPT%20to%20the,really%20started%20to%20pick%20up.

Borges, A. F. S., Laurindo, F. J. B., Spínola, M. M., Gonçalves, R. F., & Mattos, C. A. (2021). The strategic use of artificial intelligence in the digital era: Systematic literature review and future research directions. *International Journal of Information Management*, 57. https://doi.org/10.1016/j.ijinfomgt.2020.102225

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809–825. https://doi.org/10.1177/0022243719851788

Cazzaniga, M., Jaumotte, F., Li, L., Melina, G., Panton, A.J., Pizzinelli, C., Rockall, E. & Tavares, M.M. (2024). Gen-AI: Artificial Intelligence and the Future of Work. *International Monetary Fund*. https://www.imf.org/en/Publications/Staff-Discussion-Notes/Issues/2024/01/14/Gen-AI-Artificial-Intelligence-and-the-Future-of-Work-542379

Dave, P. (2024, July 31). Google Cracks Down on Explicit Deepfakes. *Wired*. https://www.wired.com/story/google-tries-to-crack-down-on-explicit-deepfakes/

Davenport T.H., & Ronanki R. (2018). Artificial intelligence for the real world. *Harvard Business Review*, January-February.

Dietvorst, B.J., Simmons, J.P. & Massey, C. (2014). Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033.

Digital Economy and Society Index (2023). *DESI 2023 dashboard for the Digital Decade* [data set].

https://digital-decade-desi.digital-strategy.ec.europa.eu/

Epoch AI (2024). Notable AI Models. https://epochai.org/data/notable-ai-models

Espinoza, J. (2024, July 15). Europe's rushed attempt to set the rules for AI. *Financial Times*. https://www.ft.com/content/6cc7847a-2fc5-4df0-b113-a435d6426c81

European Commission (2024, April 5). EU-U.S. Terminology and Taxonomy for Artificial Intelligence - Second Edition. *European Commission*. https://digital-strategy.ec.europa.eu/en/library/eu-us-terminology-and-taxonomy-artificial-intelligence-second-edition

European Commission (2024a). AI Act enters into force. https://commission.europa.eu/news/ai-act-enters-force-2024-08-01_en

European Commission (2024b). European Artificial Intelligence Act comes into force. https://ec.europa.eu/commission/presscorner/detail/en/IP_24_4123

France 24 (2024, November 3). *Proliferating 'news' sites spew AI-generated fake stories*.

https://www.france24.com/en/live-news/20240311-proliferating-news-sites-spew-ai-generated-fake-stories

Heaven, W. D. (2024). What is AI? MIT Technology Review. https://www-technologyreview-com.cdn.ampproject.org/c/s/www.technologyreview.com/2024/07/10/1094475/what-is-artificial-intelligence-ai-definitive-guide/amp/

Hermann, E. & Puntoni, S. (2024). Artificial intelligence and consumer behavior: From predictive to generative AI. *Journal of Business Research*, 180. https://doi.org/10.1016/j.jbusres.2024.114720

Hill, K. & Hsu, T. (2024, June 10). It Looked Like a Reliable News Site. It Was an A.I. Chop Shop. *The New York Times*. https://www.nytimes.com/2024/06/06/technology/bnn-breaking-ai-generated-news.html

Hurst, L. (2023, May 23). How a fake image of a Pentagon explosion shared on Twitter caused a real dip on Wall Street. *Euronews*. https://www.euronews.com/next/2023/05/23/fake-news-about-an-explosion-at-the-pentagon-spreads-on-verified-accounts-on-twitter

Iowa Environmental Council. (2024). *Drought & Iowa's Drinking Water*. https://www.iaenvironment.org/webres/file/droughtpaper_6_18_2024.pdf

Issa, H., Jaber, J., & Lakkis, H. (2024). Navigating AI unpredictability: Exploring technostress in AI-powered healthcare systems. *Technological Forecasting and Social Change*, 202, https://doi.org/10.1016/j.techfore.2024.123311

Krasodomski, A. (ed.) et al. (2024). Artificial intelligence and the challenge for global governance: Nine essays on achieving responsible AI. Research Paper. *London: Royal Institute of International Affairs*, https://doi.org/10.55317/9781784136086.

Luccioni, A.S., Jernite, Y. & Strubell, E. (2024). Power Hungry Processing: Watts Driving the Cost of AI Deployment? *ACM FAccT '24*, June 3–6, 2024, Rio de Janeiro, Brazil.

Mahmud, H., Islam, A.K.M.N., Ahmed, S.I. & Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*, 175. https://doi.org/10.1016/j.techfore.2021.121390

Maguire, G. (2024, July 26). The top US states for renewable power generation capacity. Reuters. https://www.reuters.com/sustainability/top-us-states-renewable-power-generation-capacity-maguire-2024-07-25/

Matulionyte, R. (2023, November 7). Researchers warn we could run out of data to train AI by 2026. What then? *The Conversation*. https://theconversation.com/researchers-warn-we-could-run-out-of-data-to-train-ai-by-2026-what-then-216741

Mazzucato, M. (2024, May 30). The ugly truth behind ChatGPT: AI is guzzling resources at planet-eating rates. *The Guardian*. https://www.theguardian.com/commentisfree/article/2024/may/30/ugly-truth-ai-chatgpt-guzzling-resources-environment

Metz, C., Kang, C., Frenkel, S., Thompson, S.A. & Grant, N. (2024, April 6). How Tech Giants Cut Corners to Harvest Data for A.I. *New York Times*. https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html

MIT (2024). AI Risk Repository. https://airisk.mit.edu/#Repository-Overview

Muldoon, J., Graham, M. & Cant, C. (2024). *Feeding the Machine: The Hidden Human Labour Powering AI*. Canongate Book. ISBN 9781837261826

Muldoon, J., Cant, C., Wu, B. & Graham, M. (2024). A typology of artificial intelligence data work. *Big Data & Society*, *11*(1). https://doi.org/10.1177/20539517241232632

Murphy, H. (2024, June 3). Retraining workers for the AI world. *Financial Times*. https://www.ft.com/content/a6cb4832-c5be-480d-911c-a9ba92c929d7

OECD AI Principles Overview (2024). AI System Lifecycle. *OECD*. https://oecd.ai/en/ai-principles

Pinchevski, A. (2023). Social media's canaries: content moderators between digital labor and mediated trauma. Media, Culture & Society, *45*(1). https://doi.org/10.1177/01634437221122226

Primage (2024). What is Primage? https://www.primageproject.eu/project/

Rahman-Jones, I. (2024, July 3). AI drives 48% increase in Google emissions. *BBC*. https://www.bbc.co.uk/news/articles/c51yvz51k2xo

Schaul, K., Verma, P. & Zakrzewski, C. (2024). See why AI detection tools can fail to catch election deepfakes. *The Washington Post*.

https://www.washingtonpost.com/technology/interactive/2024/ai-detection-tools-accuracy-deepfakes-election-2024/

Selwyn, N. (2024). On the Limits of Artificial Intelligence (AI) in Education. Nordisk tidsskrift for pedagogikk og kritikk, 10, 3-14.

Sippy, T., Enock, F.E., Bright, J. & Margetts, H.Z. (2024). Behind the Deepfake: 8% Create; 90% Concerned. *The Alan Turing Institute*. https://www.turing.ac.uk/sites/default/files/2024-07/behind_the_deepfake_full_publication.pdf

Snaith, B. (2023, December 22). What do we mean by "without data, there is no AI"? *The Open Data Institute*. https://theodi.org/news-and-events/blog/what-do-we-mean-by-without-data-there-is-no-ai/

Solanki, P., Grundy, J. & Hussain, W. (2023). Operationalising ethics in artificial intelligence for healthcare: a framework for AI developers. *AI and Ethics*, 3, 223-240. https://doi.org/10.1007/s43681-022-00195-z

Stephany, F. & Teutloff, O. (2024). What is the price of a skill? The value of complementarity. *Research Policy*, 53, https://doi.org/10.1016/j.respol.2023.104898

Straw, I. & Wu, H. (2022). Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction. *BMJ Health & Care Informatics*, *29*(1). 10.1136/bmjhci-2021-100457

Taylor, B.L. (2023, November 15). Long hours and low wages: the human labour powering AI's development. *The Conversation*. https://theconversation.com/long-hours-and-low-wages-the-human-labour-powering-ais-development-217038

Triguero, I., Molina, D., Poyatos, J., Del Ser, J., & Herrera, F. (2023). General Purpose Artificial Intelligence Systems (GPAIS): Properties, definition, taxonomy, open challenges and implications. *ArXiv*. https://doi.org/10.48550/arXiv.2307.14283

Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L. & Hobbhahn, M. (2024). Will we run out of data? Limits of LLM scaling based on human-generated data. Cornell University. arXiv:2211.04325

Young, J. (2024, March 24) Why AI Is So Thirsty: Data Centers Use Massive Amounts of Water. *Newsweek*. https://www.newsweek.com/why-ai-so-thirsty-data-centers-use-massive-amounts-water-1882374