



ALFIE

ASSESSMENT OF LEARNING TECHNOLOGIES AND FRAMEWORKS
FOR INTELLIGENT AND ETHICAL AI

D2.1 Initial modern AI deployments and ethical gaps

Deliverable No:	D2.1
Deliverable Name:	Initial modern AI deployments and ethical gaps
Work Package:	WP2
Task:	T2.1, T2.2
Dissemination Level:	PU
Deliverable Type:	R
Lead Organization:	KInIT
Submission month:	M12
Date:	30 September 2025



Funded by
the European Union

Funded by the European Union under Grant Agreement 101177912. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.



Project description

Acronym	ALFIE
Title	Assessment of Learning technologies and Frameworks for Intelligent and Ethical AI
Coordinator	CATALINK
Reference	101177912
Type	Research and Innovation Actions (RIA)
Programme	Horizon Europe (HORIZON)
Topic	HORIZON-CL2-2024-TRANSFORMATIONS-01-06 Beyond the horizon: A human-friendly deployment of artificial intelligence and related technologies
Start	01.10.2024
Duration	36 months
Website	alfie-project.eu
Consortium	Catalink Limited (CTL), Cyprus, Coordinator University of Brighton (UoB), United Kingdom Centre for Research & Technology (CERTH), Greece Distributed Analytics Solutions LTD (DAS), United Kingdom Edge Hill University (EHU), United Kingdom Kempelen Institute of Intelligent Technologies (KINIT), Slovakia Universitat Autònoma de Barcelona (UAB), Spain Robert Bosch Espana SLU (Bosch), Spain Eindhoven University of Technology (TU/e), Netherlands Diadikasia Business Consulting (DBC), Greece

Document history

Version	Date	Beneficiary	Description
0.1	06.11.2024	CTL	Structure proposal and initial draft
0.2	13.11.2024	CTL	Created initial table of contents
0.3	11.06.2025	CTL	Added the first section as an introduction to the literature review
0.3.1	12.06.2025	CTL	Added use case section ethical AI tools for connected automated vehicles (CAVs)
0.4	20.06.2025	CERTH	Added section explaining AI models to mitigate bias and ethical concerns
0.5	04.07.2025	UOB	Added section ethical considerations and current AI legislation
0.6	05.07.2025	KinIT	Added section use of speech-to-text and NLP in AI and its ethical considerations
0.7	07.07.2025	Bosch	Added the content for the use case compliance screening
0.8	11.07.2025	UAB	Added use case section accessibility checker for blind visitors of website
0.9	05.08.2025	EHU	Added the content for section exploring bias and ethics in knowledge graph generation
1.0	08.08.2025	TU/e	Added the content for section leveraging code generation to generate ethical and unbiased codebases
1.0.1	13.08.2025	CTL	Final editorial corrections and formatting before review
1.0.2	15.09.2025	DAS	Reviewed the deliverable and proposed amendments
1.0.3	15.09.2025	EHU	Reviewed the deliverable and proposed amendments
1.1	29.09.2025	CTL	Final version

Authors

Partner	Name(s)
KINIT	Martin Tamajka, Juraj Podroužek, Lucia Kobzova, Sebastian Kula, Adrian Gavornik
CTL	Zenon Lamprou, Haris Shekeris, Kostas Avgerinakis, Maria-Eleni Skarkala
UAB	Pilar Orero, Chiara Gunella, Sarah Anne McDonagh
UoB	Philip Haynes
DBC	Stella Theologidou, Daphne Giakoumaki
TUe	Subhaditya Mukherjee

Contributors

Partner	Contribution type	Name
EHU	Review	Nonso Nnamoko
DAS	Review	Charlie Gadd, Stelios Kapetanakis

Glossary

Acronym	Definition
AAR	Aggregate Accessibility Rating
AES	Automated Essay Scoring
AGI	Artificial General Intelligence
AI	Artificial intelligence
ANN	Artificial Neural Network
ASR	Automatic Speech Recognition
AST	Automatic Speech Translation
BMS	Building management system
CAVs	Connected Automated Vehicles
CNNs	Convolutional Neural Networks
DL	Deep Learning
ECG	Electrocardiogram
EEG	Electroencephalogram
EEO	Electrooculogram
ETD-Hub	EthiTech Dialogue Hub
FAT	Fairness, Accountability and Transparency
GA	Genetic algorithm

GDPR	General Data Protection Regulation
GenAI	Generative Artificial intelligence
GHG	Greenhouse gas
GNNs	Graph Neural Networks
GPAI	General-Purpose AI
GPTs	Generative Pretrained Transformers
HITL	Human-in-the-loop
HMS	Health monitoring system
IIoT	Industrial Internet of Things
IoT	Internet of Things
ITSs	Intelligent Tutoring Systems
JOIST	JOint Speech and Text Streaming
LIME	Local Interpretable Model-Agnostic Explanations
LLMs	Large Language Models
LSTM	Long short-term memory
LSTM – NN	Long Short-Term Memory Neural Network
MaaS	Mobility-as-a-Service
ML	Machine learning
MLM	Masked language modeling
MOOCs	Massive Open Online Courses
MT	Machine Translation
MWER	Minimum Word Error Rate
NAO	North Atlantic Oscillation
NER	Named entity recognition
NGOs	Non-government organisations
NLP	Natural Language Processing
OWL	Web Ontology Language
PAVs	Personal Aerial Vehicles
R-GCNs	Relational Graph Convolutional Networks
RDF	Resource Description Framework
RNNs	Recurrent Neural Networks
SAE	Society of Automotive Engineers
SHAP	SHapley Additive exPlanations
SLMs	Small Language Models
SVM	Support Vector Machines

TCN	Temporal Convolution Network
TFP	Total factor productivity
UAVs	Unmanned Aerial Vehicles
V2I	Vehicle-to-infrastructure
V2V	Vehicle-to-vehicle
W3C	World Wide Web Consortium
WCAG	Web Content Accessibility Guidelines
XAI	eXplainable Artificial intelligence
XGB	Extreme gradient-boosted trees

Executive Summary

This deliverable includes a comprehensive literature review that covers academic studies, international reports, and analyses related to Artificial Intelligence (AI) technologies. It explores a broad range of sectors to encompass the diverse applications of AI, with the goal of identifying and evaluating successful AI implementations and current policy developments.

The identification of AI applications and their economic and societal impacts is examined, highlighting advancements, efficiency improvements, and transformative changes across various sectors. Furthermore, it assesses the capabilities and limitations of current AI tools, with a focus on technical aspects such as Machine Learning (ML) algorithms and AI models.

The primary aim of this deliverable is to extract key lessons from successful AI deployments, illuminating the factors that contributed to their success while providing insights to guide future AI initiatives. Concurrently, it addresses the challenges faced during these deployments, including concerns related to bias, ethics, technical obstacles and societal implications, offering a comprehensive understanding of the challenges encountered in real-world AI applications within an ethical framework.

The insights gained from the current deliverable will be integrated into the EthiTech Dialogue Hub (ETD-Hub), a discussion forum that promotes conversations on AI's legal and ethical issues among AI experts, policymakers, legal professionals, and citizens, and helps shape policies and models for the ALFIE's AutoML platform.

Moreover, the deliverable includes a comprehensive review of ALFIE's trial use cases to identify existing bias and fairness issues. It identifies state-of-the-art practices currently deployed within each use case and highlights biases inherent in these approaches. This extensive research establishes a foundational understanding of the problems ALFIE aims to address within these use cases and delineates the primary areas of focus.



Contents

Executive Summary	7
1 Introduction	13
1.1 Overview	13
1.2 Structure of the deliverable	13
2 Overview of modern AI and related technologies	14
2.1 Major subfields of AI	14
2.1.1 Machine Learning	14
2.1.2 Computer Vision	14
2.1.3 Natural Language Processing	15
2.2 Current trends and advancements in different sectors	15
2.2.1 AI in medicine	15
2.2.1.1 Clinical decision making	16
2.2.1.2 Medical imaging	16
2.2.1.3 Monitoring and patient care	17
2.2.2 AI in education	17
2.2.2.1 AI in student assessment	19
2.2.3 AI in the fight against climate change	19
2.2.3.1 AI in the domain of climate modelling	19
2.2.3.2 AI in the domain of environmental monitoring	20
2.2.3.3 AI applications in building energy systems	20
2.2.3.4 AI's role in reducing transport greenhouse gas emissions (GHG)	21
3 AI in industry	22
3.1 Industrial applications of AI	24
3.1.1 Energy	24
3.1.2 Logistics	25
3.1.3 Financial services	26
3.1.4 Manufacturing	26
3.2 State-of-the-art AI tools utilised in industry	27
3.3 Legal aspects for utilizing AI at industrial level	28
3.4 Ethical considerations and challenges on using AI in industry	30
3.4.1 Transparency and explainability	30
3.4.2 Fairness and non-discrimination	31
3.4.3 Privacy and data protection	31
3.4.4 Human oversight and autonomy	32



3.4.5	Responsibility and accountability.....	32
3.4.6	Environmental impact and global inequality	33
4	Impact and challenges of AI deployments	34
4.1	Economic impacts of AI deployments.....	34
4.2	Societal and ethical impacts of AI deployments.....	35
4.3	Limitations and challenges of current deployments.....	36
5	Explaining AI models to mitigate bias and ethical concerns	39
5.1	Current technologies for explainable AI	39
5.2	The role of explainability in AI.....	41
5.3	Ethical challenges and bias in explainable AI	41
5.4	Ethical consideration and gap analysis.....	42
5.5	Future directions for resolving challenges and ethical problems	43
6	Use of speech-to-text and NLP in AI and its ethical considerations	44
6.1	State-of-the-art NLP models	44
6.2	State-of-the-art speech-to-text models	45
6.3	Challenges in speech recognition and NLP systems.....	47
6.4	Identifying ethical concerns and gaps when interacting with a system through natural language and speech	48
6.5	Mitigation strategies for ethical concerns and future directions	51
7	Exploring bias and ethics in knowledge graph generation.....	53
7.1	State-of-the-art tool for creating knowledge graphs.....	54
7.1.1	Traditional database and graph management systems.....	54
7.1.2	The emergence of graph databases.....	55
7.1.3	AI-driven tools and NLP integration.....	55
7.1.4	Future directions and challenges.....	56
7.2	Challenges in bias and ethical consideration in knowledge graphs.....	56
7.2.1	Data bias and representation bias.....	57
7.2.2	Ethical and privacy concerns.....	57
7.2.3	Lack of transparency and accountability	58
7.2.4	Directions for addressing bias and ethics.....	58
7.3	Mitigation strategies for limiting bias in knowledge graphs	58
7.3.1	Bias Detection and Correction.....	58
7.3.2	Data diversification and augmentation	59
7.3.3	Incorporating ethical frameworks.....	59
7.3.4	Human-In-The-Loop (HITL) approaches	59



7.3.5	Privacy-preserving techniques	59
7.3.6	Future directions	60
8	Leveraging code generation to generate ethical and unbiased codebases	61
8.1	Overview of code generation technologies	61
8.2	Ethical frameworks and challenges in code generation - Mitigation strategies	61
8.3	Privacy concerns in code generation	62
8.4	Future directions and recommendations	63
9	ALFIE's Pilot Use cases	65
9.1	Ethical AI tools for Connected Automated Vehicles (CAVs)	65
9.1.1	State-of-the-art algorithms for CAVs	65
9.1.2	Drowsiness Detection Techniques	66
9.1.3	Challenges and fairness in CAV systems	67
9.1.3.1	Technical challenges	67
9.1.3.2	Fairness issues	67
9.1.4	Ethical and bias considerations in CAVs	69
9.1.5	Bias and ethical considerations on emotion recognition on CAVs	69
9.1.6	Transparency and explainability	70
9.1.7	Future directions to bridge ethical gaps in CAVs domain	71
9.2	Accessibility checker for blind visitors of website	72
9.2.1	State of the art for accessibility and AI	73
9.2.2	Challenges of accessible AI and bias	74
9.2.3	Ethical considerations	74
9.2.4	Future directions	75
9.3	Compliance screening for partners	75
9.3.1	State-of-the-art algorithms for compliance monitoring	75
9.3.2	Challenges and fairness in compliance systems	76
9.3.3	Ethical and bias considerations in compliance screening	77
9.3.4	Transparency and explainability in compliance AI	77
9.3.5	Future directions to bridge ethical gaps in compliance domain	78
9.4	Identified ethical and bias concerns	79
9.5	Future directions to bridge ethical gaps	80
10	Ethical considerations	81
10.1	Current AI legislation	81
10.2	The process of implementation	83
10.3	Risk	84



10.4 Technical documentation and standards..... 85

10.5 Multi-agency cooperation 85

10.6 Critiques of the EU AI Act..... 86

10.7 Conclusion 88

11 Conclusions..... 89

References 91



Figures

Figure 1. Average percentage of enterprises employing at least one AI technology in EU Member States 2021-2023 -2024..... 23

Figure 2. Enterprises using AI technologies by size class, EU, 2023 and 2024 23

1 Introduction

1.1 Overview

This deliverable undertakes a comprehensive literature review, which covers academic research along with international reports and an analysis of modern AI and related technologies, ingesting these insights into the EthiTech Dialogue Hub (ETD-Hub), the ALFIE project's forum for dialogue with a view on formulating policy recommendations and fostering collaboration on ethical AI practices.

The examination spans diverse sectors to capture the full breadth of AI applications. Moreover, the deliverable identifies and analyses AI deployments and current policy developments and explores the economic and societal impacts of successful AI deployments. In doing so, it highlights advancements, efficiency improvements and transformative changes within various sectors. A dedicated section delves into the industrial applications of AI, particularly within the context of Industry 4.0 and Industry 5.0, showcasing how AI is transforming operations, while acknowledging the persistent challenges related to scalability, workforce integration and governance complexities. The deliverable also highlights both economic and societal impacts of AI deployments and presents the limitations and challenges of current AI deployments.

The capabilities and limitations of current AI tools are reviewed, scrutinizing technical aspects such as ML algorithms and AI models. This analysis will be conducted with a dual perspective, considering both the overall state of the art in AI technologies and the debate promoted by the ETD-Hub which will be presented in deliverable D2.2.

The overarching goal of this deliverable is to distil key lessons learned from successful AI deployments, shedding light on factors contributing to their success and extracting insights to inform future AI implementations. Simultaneously, the deliverable scrutinizes challenges encountered during AI deployments, including issues related to bias, ethical considerations, technical hurdles and societal implications. Finally, this deliverable aims to provide a nuanced understanding of the obstacles faced in real-world AI applications within an ethical framework.

1.2 Structure of the deliverable

The deliverable is divided into several main sections, presenting the current state of AI research and addressing specific aspects of modern AI deployments and their ethical gaps. Section 2 examines key subfields along with trends and sector-specific applications. Section 3 explores applications in areas, such as energy, logistics, finance and manufacturing, discussing tools, legal aspects and ethics. Section 4 presents the societal impacts, limitations and challenges of AI deployments. Section 5 reviews technologies for explainable AI (XAI), their ethical implications and future directions and examines how XAI may mitigate bias and ethical concerns. Section 6 covers the use of speech-to-text and NLP in AI and its ethical considerations. Section 7 examines bias challenges and ethics in knowledge graph generation. Section 8 considers leveraging code generation to create ethical and unbiased codebases. Section 9 presents ALFIE's trial use cases, detailing algorithms, challenges and ethical considerations. Section 10 addresses the ethical considerations and relevant AI legislation. Section 11 summarizes the key findings and concludes the deliverable.



2 Overview of modern AI and related technologies

Although AI may seem to have become widespread only in recent years, the concept dates back much further. The term AI was first coined by John McCarthy in 1956, but its mathematical foundations emerged earlier. More specifically, Alan Turing proposed a mathematical model of an ideal computer in 1936, McCulloch and Pitts introduced the first artificial neural network model in 1943 and Donald Hebb outlined the earliest idea of machine learning in 1949 (Jiang et al., 2022, Toosi et al., 2021). From 2000 onwards, there has been a flourishing of AI research and various breakthroughs, especially due to advances in Machine Learning (ML) and Deep Learning (DL) (Jiang et al., 2022).

The following sections present a comprehensive overview of modern AI deployments and their associated ethical gaps and the current trends and advancements of AI in various sectors.

2.1 Major subfields of AI

2.1.1 Machine Learning

The first subfield of AI to be reviewed here is machine learning (ML). ML is, broadly, the field of study and research in which statistical algorithms are developed which, together with optimization methods, enable computers to recognise patterns in large datasets and then detect and apply these patterns when encountering new, previously unseen datasets. Different techniques of machine learning have been developed, depending on the degree in which human input is required and the amount and kind of data used. The key techniques are supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning and deep learning¹. Each technique uses its own set of algorithms, which enable learning to take place.

Deep learning is a key subdomain of machine learning which has produced significant advances in recent years. Deep learning algorithms are based on neural networks with many layers, in a process loosely based on the way the human brain purportedly works. Due to their powerful network architectures, deep learning models enable the processing of significantly larger amounts of data than other machine learning models. Due to its predictive and inference power, deep learning has found many applications in other subfields of AI such as natural language processing and computer vision, as well as in other scientific disciplines such as medical imaging analysis, climate modelling and bioinformatics among others (I. Mienye & Swart, 2024).

2.1.2 Computer Vision

Another subfield of AI is computer vision. If machine learning, in simple terms, allows machines to learn as humans do, then computer vision is the field of research aiming to make computers see the way humans do. In other words, it is about enabling computers to discern information and make inferences from visual stimuli much in the same way humans do. At present, due to advances in the field, computers may extract information from data sources such as sets of still images, video images including from various cameras and medical images.

¹ <https://ischoolonline.berkeley.edu/blog/what-is-machine-learning/>



The backbone of computer vision, in terms of computational arsenal, is deep learning together with convolutional neural networks (CNNs) (Zhao et al., 2024). CNNs are a special class of neural networks which are used in image classification and object recognition tasks². Deep learning algorithms have significantly aided progress in computer vision tasks such as optical flow and segmentation, whilst CNNs play an important role in tasks such as image classification.

The most standard functions performed by computer vision systems are image acquisition, pre-processing, feature extraction, segmentation, high-level processing and decision making.

Computer vision systems have a very broad range of applications. For instance, they are used in biomedical imaging, in driver drowsiness detection, in surveillance, in species identification systems, autonomous vehicles and more general navigation, in monitoring agricultural crops and many more applications across various domains of industrial application and cutting-edge scientific research.

2.1.3 Natural Language Processing

Natural Language Processing (NLP) is another subfield of AI, relevant to the ALFIE project. NLP is the subfield of AI whose aim is to enable humans and computers to communicate using natural (human) language. The three approaches to NLP are the symbolic one, the statistical one and the deep learning one. The symbolic approach to NLP was historically the first one to be researched and as such its scalability and functionality are quite limited judged from the perspective of the state of the art. It relied on rules and gave answers only to specific prompts. Both the statistical approach and the deep learning one are machine learning approaches and as such have significantly more capabilities. The statistical approach, which relies on decision trees but most successfully on hidden Markov models, enabled the introduction of spellcheckers and T9 – text on nine keys, used in phones³. Finally, deep learning models have given a boost to NLP research and applications, as they allow the use of enormous volumes of raw, unstructured data and voice in their training to become ever more accurate (H. Liu, 2025). The key NLP tasks are text and speech processing, morphological analysis, syntactic analysis, lexical semantics, relational semantics, discourse and higher-level NLP applications. A special advance in NLP which has transformed everyday experience in the last 3 years has been the development of Large Language Models (LLMs), with the most prominent ones being Generative Pretrained Transformers (GPTs). LLMs are neural networks with billions of parameters, trained on vast datasets of unlabelled text (Bi et al., 2024). LLM models such as GPT-4o and Deepseek have displayed exceptional capacities in the performance of reasoning and cognitive tasks and have in certain cases performed better than humans (Shultz et al., 2025).

2.2 Current trends and advancements in different sectors

2.2.1 AI in medicine

Medicine and more generally healthcare is a sector which has seen significant benefits from the use of AI. Healthcare has seen an explosion of AI-powered medical applications and other AI-powered breakthroughs in biomedicine such as designer drugs and advances in diagnosis

² <https://www.ibm.com/think/topics/convolutional-neural-networks>

³ <https://www.ibm.com/think/topics/natural-language-processing>

through the use of machine learning in imaging and the facilitation of patient experience. This explosion and more generally the success of AI in biomedicine and healthcare may play a significant role in ensuring the permanent incorporation of AI in human life and in ensuring that there are no more AI nuclear winters. This is because medicine is a high-risk but also highly rewarding sector. Any potential major setbacks of AI in medicine may be fatal to it, given the cost in human lives and wellbeing. However, on the other hand, any major successes will facilitate acceptance and social uptake of AI technologies and mitigate any legal, ethical and philosophical concerns.

So far and more specifically, AI has been instrumental in bringing about radical change in aspects of healthcare such as clinical decision making, hospital operations and management, medical imaging and diagnosis, as well as monitoring and patient care (Maleki Varnosfaderani & Forouzanfar, 2024). In the following sections, each of these aspects will be expanded further.

2.2.1.1 Clinical decision making

AI has enabled significant advances in clinical diagnosis and prognosis, as well as in personalized medicine. Cardiology is a discipline which has benefited the most out of the introduction of AI in medicine (Karatzia et al., 2022). Significant results have been achieved for example in the identification of arrhythmia using a support vector machine and electrocardiogram data (Dilsizian & Siegel, 2018), the prediction of coronary artery disease using Cardiac Computed Tomography Angiography and clinical data through a boosted ensemble algorithm (Lu et al., 2022) and the prediction of early coronary revascularisation within 90 days after single proton emission computerised tomography (SPECT) myocardial perfusion imaging (MPI), based on clinical and SPECT data and using an Ensemble LogitBoost algorithm (Koulaouzidis et al., 2022). In pharmacology, Graph Neural Networks (GNNs) have provided useful insights into the resolution of the Drug-Drug Interaction (DDI) problem (X. Lin et al., 2020).

AI approaches have also helped progress in personalized medicine. More specifically, in drug development, the ML kernel-based regression algorithm which consisted in an ensemble model of 440 CGKronRLS regressors, had the best performance in a Drug-Kinase Binding Prediction Challenge (Cichońska et al., 2021), whilst a Long Short-Term Memory Neural Network (LSTM – NN) has been used for screening and affinity predictions (Chakravarti & Alla, 2019). In the advanced management of Type 1 Diabetes, several AI techniques, such as extreme gradient-boosted trees (XGB), support vector classifiers and individualized neural networks are used to predict specific events such as hypoglycemia and Temporal Convolution Network (TCN) is used to benchmark glucose prediction algorithms, among other algorithms for other tasks (Vettoretti et al., 2020).

2.2.1.2 Medical imaging

Medical imaging has been playing a very important role in medicine ever since the first iconic medical X-ray of Wilhelm Röntgen's wife's hand in 1895. At present, many modalities of medical imaging exist, with some key techniques being X-Ray Radiography, X-Ray Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Ultrasonography (ultrasound), Elastography, Optical Imaging, Radionuclide Imaging including Positron Emission Tomography (PET) and SPECT, Infrared Thermography and Terahertz Imaging (Kasban et al., 2015). A medical discipline which has benefited from such advances is radiology, where already from 2019 a deep learning model based on a 121-layer convolutional neural network

(CheXNet) outperformed a group of board certified diagnostic radiologists (Patel et al., 2019). In general, Convolutional Neural Networks (CNNs) are mostly used in medical imaging, often with excellent results, such as Deeplab v3+ which has recorded an accuracy of 95,76% in the segmentation of gastric cancer (J. Wang & Liu, 2021). Pathology has similarly seen rapid advances, with the recent establishment of the field of digital pathology which has benefited significantly from the development of whole-slide imaging (Tsaris et al., 2023) and the use of deep learning algorithms. Ophthalmology is another medical discipline which has significantly benefited from advances in AI. For example, the deep learning model system DeepDR Plus has successfully predicted the progression and has personalised the screening intervals in a period of five years in patients at risk of developing diabetic retinopathy, which is the leading cause of preventable blindness (Dai et al., 2024).

2.2.1.3 Monitoring and patient care

AI-powered wearable devices provide continuous monitoring and data collection, which may be used for a variety of diagnostic and monitoring tasks. A classification of such devices would include categories such as wrist-worn, head mounted, ornament, e-textiles, e-patches and sports and fitness devices (Nahavandi et al., 2021). Various devices such as wrist-watches, smart T-shirts and e-tattoos are specific examples of wearable devices. Some challenges of using wearables are obtaining good quality data in a secure manner, the precision of the device's localization and the lack of diverse and representative datasets, among others (Nahavandi et al., 2021; Sreeharsha et al., 2024)

An example of the use of wearable sensors in biomedicine is the estimation of Parkinson's Disease-associated tremors during free body movement (Hssayeni et al., 2019). Using an ensemble model based on gradient tree boosting the authors achieved a high correlation ($r = 0,96$ using held-out testing) between the estimated and clinically assessed tremor subscores.

Another example of a successful use of an AI-powered wearable device is the use of an architecture combining wavelet transforms and multiple LSTM recurrent neural networks in order to continuously monitor heart rate and thus serve in the prevention of deaths from cardiovascular diseases (Saadatnejad et al., 2020). According to the authors and corroborated by Nahavandi et al., the device achieves both a very accurate electrocardiogram (ECG) classification and a low power consumption, which is an added advantage.

Furthermore, a broad variety of AI Voice Assistants (VAs) has been developed with the aim of providing instructions, guidelines and navigation to patients. For example, CardioCube, an Electronic Health Record (EHR) - integrated VA, has automated tasks such as medical data collection, indexing and documentation with an accuracy-rate of 97,51% between verbally provided information and the corresponding EHR database (Jadczyk et al., 2019). Voice assistants also provided innovative solutions to various problems during the COVID-19 pandemic (Jadczyk et al., 2021).

2.2.2 AI in education

Education, be it primary, secondary, higher or lifelong, is another high-stakes sector where AI is being used with a significant degree of success. AI has the potential to revolutionise the education sector, however it also faces significant ethical and other challenges. Potential benefits could be the improvement of students' learning outcomes, especially given the possibility of personalized learning, something which will significantly benefit underserved populations. Challenges include data privacy and security; the importance to maintain a

healthy student-teacher relationship; the exposure of children to misinformation; and the possibility of bias and discrimination, as due to the fact that the datasets used to train the AI models may not be targeted specifically to children (Kamalov et al., 2023).

A key set of AI tools used in student learning are Intelligent Tutoring Systems (ITSs). Research in ITSs began already from 1970 when a first set of computer programs named SCHOLAR and designed for education was written (Carbonell, 1970). ITSs may model each student's psychological states, such as motivation and emotions, their prior knowledge and skills, as well as track their progress, provide feedback and select appropriate problems for the students to practice with (C. Lin et al., 2023). An example of an ITS is SeisTutor, a system which aims to provide a custom-tailored tutoring strategy and mimic the cognitive intelligence of a human tutor in the delivery of a specific course curriculum (Singh et al., 2022). SeisTutor uses a CNN-based approach to capture the learner's emotions through a webcam in order to guide the instructor's future actions. It also uses dictionary-based sentimental analysis in order to gauge the learner's overall understanding. When it comes to the evaluation of the tool, 87% of the learners agreed that it improved their learning performance and outcomes, whilst most of them appreciated the AI elements of the tool (Singh et al., 2022).

AI has also significantly affected teaching, besides learning. Chiu et al., (2023) have identified three roles which AI may play in teaching: a) it may provide adaptive teaching strategies, b) it may enhance the ability of teachers to teach and c) it may support their professional development.

Yang et al., (2020) evaluated the employment of an educational robot. As per the authors, the educational robot is composed of three layers, the motion layer which controls the walking and speed of the robot, the motion management layer which controls the overall motion state of the robot and the robot's microcomputer which controls the communication with the students through the voice and action recognition system. The model governing the voice interaction between the robot and the student is composed of three modules: a) the speech recognition module, b) the interaction management module and c) the speech synthesis module. The model used for the speech recognition module consists of a deep neural network-based acoustic model which uses unsupervised learning for the initial assignment of weights to learning links and then uses supervised learning to finetune the network's parameters.

The second component of the teaching tool, besides the robot, is the construction of a hybrid teaching course in Physical Education. With the use of the robot, the teacher is enabled to record videos of the set exercises whilst using features such as slow motion and verbal description to walk the students through technical parts. Furthermore, the students are able to submit coursework via the online teaching platform and the teachers may provide them with good-quality feedback. The teaching process is broken into three stages, the autonomous learning before class stage, the classroom practice teaching stage and finally the after-class consolidation stage. It is worth remarking that the authors stress that through this mode of teaching, the traditional "teaching before learning" mode of classroom teaching is transformed through hybrid teaching into a new mode, that of "learning before teaching" (Yang et al., 2020).

In evaluating the proposed robot and hybrid teaching method, the authors found that the voice interaction module had a success rate of more than 90% in speech recognition, however this result was affected by noise and by non-standard uses of language. Furthermore, the researchers detected a higher interest of students in sports, measured before and after teaching the course, when the robot is involved compared to traditional teaching. Students'



attitudes towards learning sports also showed a bigger improvement when the hybrid method was used. However, the authors stress the small size of the sample and point out that larger-scale studies are required to establish the overall advantage of using robots to all students of physical education.

2.2.2.1 AI in student assessment

AI tools may also play two roles in student assessment: a) they may aid in automating marking and b) they may aid in predicting student performance, especially in teacherless online courses such as Massive Open Online Courses (MOOCs). The authors (Kumar & Boulanger, 2020) present an example of the application of explainable AI (XAI) to an Automated Essay Scoring (AES) software program. They describe an AES system which a) predicts both holistic and specific rubric scores b) uses deep learning, more specifically, multi-layer perceptron neural networks, to predict writing quality and c) uses SHapley Additive exPlanations (SHAP) to explain the decision-making process behind the predictions. The most significant findings of the study were that a) the more hidden layers were added to the neural network, the more the descriptive accuracy of the SHAP explanation increased, hence the AES system becomes more trustworthy; b) faster SHAP implementations, such as DeepSHAP and GradientSHAP, are as accurate as slower ones, thus enabling the provision of real-time feedback to students and d) SHAP may be used to provide personalised, formative and fine-grained feedback to students and even forecast the improvement of a student's score if they follow the feedback.

2.2.3 AI in the fight against climate change

Climate change is one of the biggest existential challenges for humanity at present. AI may play a significant dual role in combating this phenomenon: a) it can help scientists understand the phenomenon better and b) it can play a significant role in the development of solutions that would mitigate or reverse its effects (Cowls et al., 2023). It should be remarked that both the use of AI in general and the use of AI in the search for solutions for climate change mitigation have political and ethical dimensions as well and hence should be viewed through that lens. The use of AI for general purposes in particular invites a special ethical dilemma as the development of AI systems and the associated research leaves behind it a significant greenhouse gases footprint.

2.2.3.1 AI in the domain of climate modelling

Models play a very important role in enabling researchers and forecasters to understand the dynamics of climate and to produce accurate predictions both for the weather (short-term, up to for example ten days) and the climate (more long-term predictions). As defined by the authors of a review on the topic (Slater et al., 2023), dynamical models use numerical modelling to solve dynamical physical processes. On the other hand, the authors define data-driven models as models which include empirical, statistical and machine learning methods ranging from simple linear regression to deep neural networks. Hybrid prediction combines the strengths of these two types of models by fusing their elements.

A first such hybrid model is described by Nevo et al., (2022). The authors describe a system consisting of the following four subsystems: a) data validation; b) stage forecasting; c) inundation modelling; and d) alert distribution. Long short-term memory (LSTM) is used in the stage forecasting system whereas flood inundation modelling is performed using thresholding and manifold models. The authors found that the LSTM model performed better than linear

models, whilst the thresholding and manifold models displayed similar performance. The model succeeded in sending out more than 100,000,000 flood alerts to affected populations, local authorities and emergency organisations (Nevo et al., 2022).

A second hybrid model was developed by researchers (Moulds et al., 2023) to predict multi-year average weather high streamflows in the UK. The researchers developed a statistical-dynamical framework which used a large multi-model ensemble of decadal hindcasts, furthermore using a lagged ensemble approach in order to increase the ensemble members. Initially, the researchers found a low skill of raw decadal predictions - this was mainly attributed to the underestimation of the North Atlantic Oscillation (NAO) climate effect. When NAO-matching was factored in, a significant improvement in the prediction skill was observed.

2.2.3.2 AI in the domain of environmental monitoring

Environmental monitoring is primarily used to understand the phenomenon of climate change, although the data derived from it may be used to generate solutions to specific climate-change induced problems. Air or water quality, soils, ice or other environmental parameters may be monitored and furthermore, monitoring may be directed not only towards the present state of the parameter monitored but it may also be used to make predictions about its future states.

Forzieri et al., 2022 monitor early warning signs of declining forest resilience by integrating satellite-based vegetation indices with machine learning. The authors used a Random Forest regression model with 100 regression trees. The trees' depth and number of predictors to sample at each node were identified using Bayesian optimization. The authors used this model to study how forest resilience changed in the period 2000-2020. From the patterns that emerge out of the analysis of a vast volume of data thanks to machine learning, the authors corroborated the existence of common large-scale climate drivers. Furthermore, they discovered that approximately 23% of intact undisturbed forests have already reached a critical threshold and are experiencing further degradation in resilience (Forzieri et al., 2022).

2.2.3.3 AI applications in building energy systems

Farzaneh et al., (2021) review a large amount of AI-based innovations in the domain of building energy systems. They stress that such innovations have significant effects in various domains such as comfort, energy, maintenance, safety and design. They also note that the priority of operationalized concepts such as the smart grid or smart buildings, as well as the building management system (BMS), is both to enhance the energy efficiency of homes but to also improve the experience of living in such buildings and homes.

A first innovation (Farzaneh et al., 2021) is the prediction of the energy consumption in building with the use of a combination of neural networks and a hybrid genetic algorithm-adaptive network-based fuzzy inference system (GA-ANFIS model), as described by (K. Li et al., 2011). The authors use a genetic algorithm (GA) and an Artificial Neural Network (ANN) and their results indicate that the use of the GA-ANFIS model yields better results than using the ANN.

A second innovation described by Farzaneh et al., (2021) has to do with building a prediction model to calculate the hourly building cooling load in an office building in Guangzhou, China. The AI algorithm used for this study was the Support Vector Machine (SVM), whilst the key results of the study, according to its authors, were that using SVM one achieves a better accuracy and generalization than the traditional back-propagation neural network model and that hence the former is more effective for building cooling load prediction (Q. Li et al., 2009).



Finally, a third innovation listed by Farzaneh et al. (2021) is the detection of abnormal functioning conditions and the generation of fault signatures for various fault types using fuzzy logic and ANNs in the novel health monitoring system (HMS) for a variable air volume unit. According to the authors (Allen et al., 2016), using this method achieved a lower cost of ownership and operations, improved efficiency of equipment and resulted in fewer failure events, reduced long-term maintenance costs and saw an improvement of the length of the asset life cycles.

2.2.3.4 AI's role in reducing transport greenhouse gas emissions (GHG)

Emissions of GHG in the transport sector constituted up to a third of global GHG emissions in 2015 (Solaymani, 2019). Machine learning algorithms such as Genetic Algorithm (GA), Support Vector Machine (SVL), Naive Bayes (NB), k-means clustering and DL algorithms such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Restricted Boltzmann Machine (RBM) are already being explored because of their potential to optimize energy efficiency, reduce emissions and enhance safety in transport (Mirindi et al., 2025). A key idea that motivates research in this domain is the nexus between the smart city, AI and transport (Nikitas et al., 2020). Innovations linked to this nexus, such as Connected Automated Vehicles (CAVs), Unmanned Aerial Vehicles (UAVs) and Personal Aerial Vehicles (PAVs) and concepts such as Mobility-as-a-Service (MaaS) hold enormous promise in revolutionizing the transportation experience in a significantly more sustainable manner. The adoption of these technologies would aid significantly in reducing congestion, improved driving efficiency, reduced energy emissions and decreased environmental pollution (Pan et al., 2024). These technologies and concepts in turn rely heavily on AI-enabled notions such as the Internet of Things, the Physical Internet and Industry 4.0. However, most of them are still at the research or early marketisation phase rather than being embedded in the industry, production and lifestyle domains.

3 AI in industry

AI is fundamentally transforming traditional industries across a wide range of operational functions and strategic domains. Its integration within sectors such as manufacturing, logistics, finance, banking, as well as energy and transport exemplifies how AI is reshaping decision-making processes, optimizing resource allocation and enabling more adaptive, data-driven ecosystems.

Industry 4.0 (I4.0) represents a transformative paradigm, emphasising the integration of advanced technologies to revolutionise products, services and processes through seamless connectivity and agile decision-making (Passalacqua et al., 2025). The Industry 4.0 paradigm saw the transition to continuous, real-time data collection from interconnected devices and systems, powered by the Internet of Things (IoT) and Big Data. Machine learning algorithms began to analyze massive data sets collected from IoT devices to detect patterns, predict failures and recommend maintenance actions (Mikołajewska et al., 2025).

Building on this foundation, the European Commission has introduced a progressive approach, termed Industry 5.0 (I5.0), marking a significant evolution to I4.0. I5.0's vision accentuates a more human-centric approach to technological design and integration in the workplace, emphasizing a balance between technological progress and human welfare (European Commission 2021).

In essence, Industry 5.0 aims to build upon the technological advances of Industry 4.0 by embedding values of human well-being, sustainability and resilience at the core of industrial evolution, ensuring that the future of AI-driven ecosystems is not only more intelligent, but also more inclusive and socially responsible.

Figure 1 presents the average proportion of enterprises with a workforce of at least ten employees across EU Member States that reported the use of at least one AI technology in the years 2021, 2023 and 2024, as derived from Eurostat data⁴. These figures encompass a broad range of economic activities, classified according to the NACE Rev. 2 framework, thereby providing a comprehensive overview of AI adoption patterns across Europe's industrial and service sectors. The indicator encompasses various AI technologies, including text mining systems, speech recognition tools, natural language generation applications such as image recognition systems, machine learning algorithms, AI-driven process and equipment automation and autonomous robotics.

⁴https://ec.europa.eu/eurostat/databrowser/view/isoc_eb_ain2__custom_17427103/default/table?lang=en

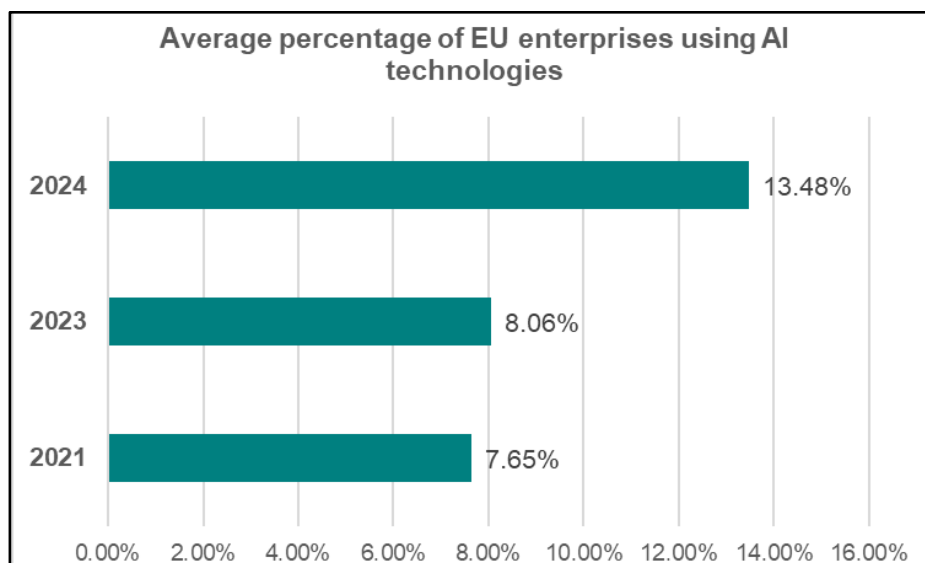


Figure 1. Average percentage of enterprises employing at least one AI technology in EU Member States 2021-2023 -2024⁵.

The data presented in Figure 1 illustrate a clear upward trend in the adoption of AI technologies by enterprises across EU Member States between 2021 and 2024. Specifically, the average share of enterprises employing at least one AI technology increased from 7,65% in 2021 to 8,06% in 2023, reaching 13,48% in 2024.

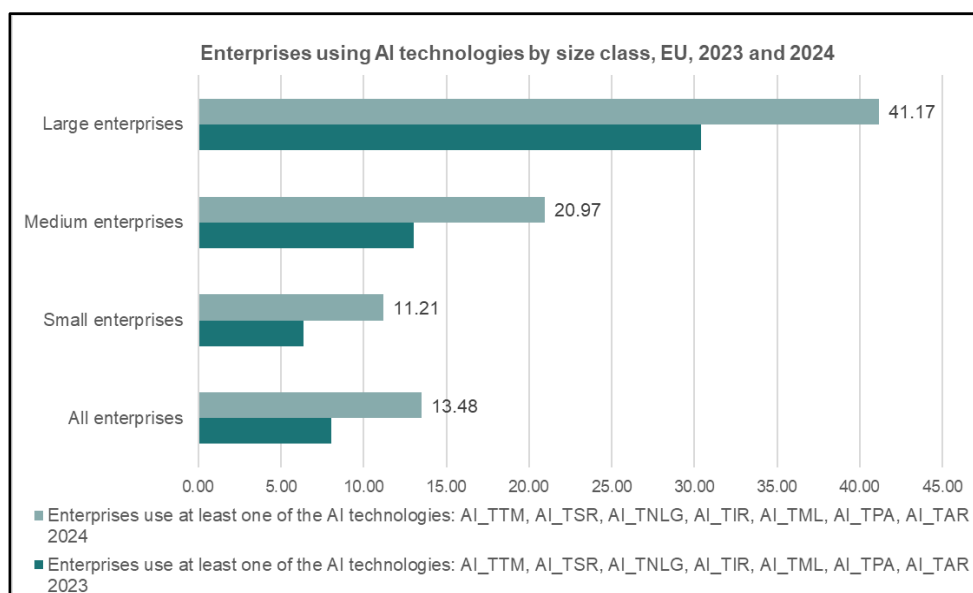


Figure 2. Enterprises using AI technologies by size class, EU, 2023 and 2024⁶

⁵ Source: Own elaboration based on Eurostat dataset “Artificial intelligence by NACE Rev. 2 activity” (isoc_eb_ain2__custom_17427103), last updated 8 April 2025

⁶ Source: Eurostat (online data code: isoc_eb_ai) https://doi.org/10.2908/ISOC_EB_AI

However, adoption is not evenly distributed across enterprises of different sizes. As shown in Figure 2, in 2024, 41,2% of large enterprises reported using at least one AI technology, compared to 21,0% of medium-sized enterprises and only 11,2% of small enterprises. This disparity highlights how resource availability, digital infrastructure, and organisational capacity strongly influence AI uptake, with larger firms positioned to scale adoption more rapidly than their smaller counterparts.

This notable rise underscores the growing integration of AI across diverse industrial and service sectors within the EU. It reflects a broader shift toward data-driven operations and advanced automation, aligning with strategic objectives under Industry 4.0 and the emerging human-centric imperatives of Industry 5.0. The figures also emphasize that while AI uptake is accelerating, significant scope remains for further diffusion of these technologies across European enterprises to achieve widespread digital transformation.

The World Economic Forum (2025)⁷ highlights that industries worldwide are now embedding AI into core operations, moving decisively beyond mere experimentation. Yet despite this momentum, approximately 74% of companies still report substantial obstacles when trying to scale AI initiatives, with challenges ranging from fragmented data infrastructure and interoperability issues to skill shortages and governance complexities.

3.1 Industrial applications of AI

A growing number of industries are now embedding AI within their operational and strategic frameworks, recognizing its potential to drive a new era of industrial performance. AI serves not merely as a technological add-on, but increasingly as a capability that enables organizations to transition from reactive to predictive modes of operation. Through developments such as digital twins that accurately replicate physical processes, traditional industrial ecosystems are becoming markedly more proactive, adaptive and intelligent. This evolution helps prevent major disruptions, reduce operational losses and sustain continuity across workflows. This section examines how various sectors are integrating AI into their core processes, focusing on key domains including energy, logistics, financial services and manufacturing.

3.1.1 Energy

The energy sector faces growing demands for efficiency, reliability and sustainability, all under increasingly complex operating conditions. AI has emerged as a critical enabler in this transformation. One of the most widespread applications of AI in this sector is the prediction of energy supply and demand, where machine learning models analyze vast datasets generated by smart meters, weather stations and grid sensors to support informed decision-making⁸. Beyond forecasting, AI is being integrated into a wide range of operational and planning processes across the energy value chain. It contributes to optimizing asset performance, enabling predictive maintenance and enhancing infrastructure resilience. Overall, AI is gradually reshaping how utilities manage both renewable and conventional energy systems.

⁷https://reports.weforum.org/docs/WEF_AI_in_Action_Beyond_Experimentation_to_Transform_Industry_2025.pdf

⁸ <https://www.iea.org/commentaries/why-ai-and-energy-are-the-new-power-couple>

Based on ENEL's report⁹ in 2024, AI is applied in wind energy through the use of smart sensors and predictive algorithms that support various maintenance scenarios, improving efficiency and boosting clean energy output. The same report also highlights how predictive algorithms streamline the logistics of wind turbine installation and enhance overall plant operations, while computer vision technologies assist in the early detection of infrastructure faults.

In a comprehensive review, Q. Wang et al. (2025) demonstrate how AI integration significantly enhances renewable energy systems by optimizing photovoltaic array configurations and improving energy production efficiency. The study also highlights AI's contribution to advancing innovation in energy materials and its alignment with enabling technologies such as blockchain and the IoT. Despite these advancements, the authors identify persistent economic, environmental and ethical challenges that underscore the need for robust governance frameworks and international cooperation.

3.1.2 Logistics

AI is being deployed extensively in logistics operations, improving sectors such as route optimization, warehouse and inventory functions and picking and handling of packages; these technologies are redefining efficiency, responsiveness and resilience. While challenges around scalability and workforce integration remain, the ongoing investments by industry leaders signal AI's central role in shaping the next-generation logistics infrastructure.

A notable example is DHL's collaboration with Dorabot to deploy "DHLBots" AI-enabled sorting robots across its logistics facilities. These robots are capable of processing over 1,000 small parcels per hour with an accuracy rate of 99%, significantly reducing the need for secondary sorting and improving overall throughput. Following a successful pilot in Miami, DHL has expanded the deployment of this technology to facilities in Asia-Pacific, achieving a 40% increase in sorting capacity¹⁰. Boston Dynamics¹¹ "Stretch" robot is another example of AI-driven automation in logistics, utilizing machine learning and computer vision to autonomously unload packages of varying shapes and sizes in warehouse settings. Complementing this, other AI-enabled systems like "Spot," designed for industrial inspection tasks, further highlight the growing role of intelligent robotics in enhancing operational efficiency and safety across logistics operations. Moreover, Amazon has deployed over one million AI-enabled robots in their fulfillment centers¹² to perform tasks such as sorting, picking and transporting packages; substantially reducing processing times and operational costs. Amazon has also introduced a new generative AI foundation model, DeepFleet, designed to make its entire fleet of robots smarter and more efficient¹³.

In summary, AI is making particularly rapid progress and is being adopted more quickly and broadly in the logistics and transportation sector, outpacing adoption rates seen in other industrial domains. AI is transforming core operations from route optimization and warehouse automation to predictive analytics and intelligent parcel handling, enabling logistics providers to respond faster to market demands while minimizing errors and costs. At the same time, this accelerated uptake also raises concerns, including potential job displacement, heightened

⁹ <https://www.enel.com/media/word-from/news/2024/07/ai-future-on-the-road-to-innovation>

¹⁰ <https://www.dhl.com/global-en/delivered/innovation/ai-in-logistics.html>

¹¹ <https://bostondynamics.com/products/stretch/> <https://bostondynamics.com/products/spot/>

¹² <https://www.aboutamazon.com/news/operations/amazon-million-robots-ai-foundation-model>

¹³ <https://www.aboutamazon.com/news/operations/amazon-million-robots-ai-foundation-model>



cyber risks from connected robotics and challenges in ensuring workforce reskilling and system resilience.

3.1.3 Financial services

AI has become an essential capability for many financial institutions, supporting areas such as customer operations, risk assessment and product development. According to McKinsey's 2025 Global AI survey¹⁴, the financial services industry ranks among the top sectors investing in AI, alongside high tech and telecommunications. Financial firms are increasingly deploying machine learning algorithms to enhance decision-making accuracy, automate credit scoring and detect fraudulent transactions in real-time. While AI enables hyper-personalized customer experiences through intelligent chatbots, recommendation systems and behavioural analytics, these tools may also amplify privacy risks and create challenges for data governance. In risk and compliance, AI supports the early identification of anomalies, improves regulatory reporting and bolsters cybersecurity frameworks, yet its deployment also introduces regulatory tensions around explainability in sensitive domains and raises risks of algorithmic bias and discrimination.

3.1.4 Manufacturing

In the emerging era of smart manufacturing, characterized by the integration of advanced technological solutions into manufacturing applications and infrastructure, artificial intelligence plays a pivotal role. AI-driven technologies facilitate substantial advancements across various manufacturing functions, enabling the sector to achieve significant operational improvements and establish new industry standards.

Among the diverse operational areas within manufacturing, the following sections focus on key functions where AI has brought especially clear benefits, along with other important ways AI is shaping the industrial sector.

- **Maintenance:** AI is incorporated in the prediction of real-time maintenance needs and the monitoring of system lifespan. Early detection of anomalies allows for timely interventions, significantly reducing equipment downtime and maintenance costs. Recent advancements such as the integration of generative AI through digital twins for predictive fault diagnosis further enhance industrial efficiency, operational cost-effectiveness and workplace safety. In predictive maintenance, Generative AI (GenAI) digital twins provide highly realistic operational scenarios, effectively identifying potential failure modes that might otherwise go undetected using traditional diagnostic methods (Mikołajewska et al., 2025). In summary, given the manufacturing sector's inherent complexity and dependence on advanced technological infrastructures, the strategic deployment of AI holds substantial potential to enhance operational efficiency, reduce costs and improve workplace safety and resilience.
- **Quality control:** AI has transformed quality assurance in manufacturing through advanced computer vision and machine learning systems that detect defects with precision and speed beyond human capability (Junior et.al., 2025). AI-driven visual inspection solutions automatically identify surface anomalies, dimensional deviations

¹⁴https://www.mckinsey.com/~media/mckinsey/business%20functions/quantumblack/our%20insights/the%20state%20of%20ai/2025/the-state-of-ai-how-organizations-are-rewiring-to-capture-value_final.pdf



and assembly inconsistencies in real time, ensuring consistent product standards and minimizing waste. For example, Siemens Inspekto system integrates ML with industrial imaging to autonomously learn expected quality characteristics and identify deviations without extensive pre-programmed rules, significantly enhancing flexibility and efficiency on production lines (Siemens, 2025¹⁵).

- **Supply chain management:** AI can keep all parts of a supply chain in balance with its ability to find patterns and relationships unlike a traditional non-AI system (IBM, 2025). By analysing extensive historical data alongside real-time inputs, AI systems improve demand forecasting and lead time estimation, supporting more accurate inventory and production planning than classical approaches allow (Plathottam et.al., 2023). In supply chain management, AI is primarily applied through ML for predictive analytics, reinforcement learning for dynamic optimization, NLP for extracting insights from textual data and computer vision for automated warehouse operations. At the same time, scaling these technologies poses challenges, including safety concerns in robotics, the explainability and robustness of AI-driven decisions and integration difficulties in large and heterogeneous supply systems.

3.2 State-of-the-art AI tools utilised in industry

According to the Eurostat dataset presented above titled “Artificial intelligence by NACE Rev. 2 activity”, the classification of AI technologies adopted by enterprises includes the following categories:

- Text mining and natural language processing
- Speech recognition and generation
- Image recognition and processing
- ML and DL for data analysis
- AI-driven workflow automation and decision support
- Autonomous robotics for physical operations

A growing number of industries are also leveraging advanced analytical AI and beginning to explore GenAI for creative tasks, simulation and conversational interfaces. While GenAI is still in its early stages of industrial adoption¹⁶, it holds significant potential for tasks such as automated design, predictive modelling and synthetic data generation.

Moreover, Explainable AI (XAI) is emerging as a critical tool to ensure transparency and trust in AI-driven decisions, particularly in high-stakes sectors like healthcare, finance and energy. It enables stakeholders to understand and interpret complex model outputs, aligning with ethical principles and regulatory requirements.

Another foundational element is the Industrial Internet of Things (IIoT), which combines AI with connected devices to enable real-time monitoring, predictive diagnostics and closed-loop control systems. When integrated with digital twins, AI models can simulate industrial processes with high accuracy, offering a virtual environment for testing, optimization and fault prediction (Mikołajewska et al., 2025).

¹⁵Industrial AI – Automated quality inspection with Inspekto. Available at: <https://www.siemens.com/global/en/products/automation/topic-areas/industrial-ai/inspekto.html>

¹⁶https://reports.weforum.org/docs/WEF_AI_in_Action_Beyond_Experimentation_to_Transform_Industry_2025.pdf

While AI deployments are expanding beyond pilot phases, challenges related to scalability, data integration and workforce readiness remain. Nonetheless, AI is increasingly viewed not just as a set of tools, but as a strategic capability essential for building agile, intelligent and sustainable industrial systems (Stanford HAI, 2025¹⁷).

3.3 Legal aspects for utilizing AI at industrial level

The use of AI in industrial settings brings significant economic and operational benefits but is also subject to substantial legal and regulatory challenges. In the EU, these challenges are primarily addressed by the AI Act¹⁸, a central component of the EU's digital strategy for trustworthy AI, together with the General Data Protection Regulation (GDPR), sector-specific legislation (e.g., Machinery Regulation, Medical Devices Regulation) and product safety laws.

The AI Act applies to both public and private actors, including providers (developers or those placing AI systems on the market) and deployers (users of AI systems), regardless of whether they are established within or outside the EU, provided that the systems are placed on the EU market, or their output is used within the Union.

It introduces a risk-based regulatory framework, classifying AI systems into four categories:

1. **Unacceptable risk:** AI systems posing a clear threat to safety, fundamental rights, or democratic values are prohibited. This includes manipulative AI, exploitation of vulnerabilities, social scoring, untargeted biometric data scraping and real-time remote biometric identification in public spaces (for law enforcement purposes).
2. **High risk:** AI systems with significant potential to affect health, safety, or fundamental rights face stringent compliance obligations. Examples include:
 - AI safety components in critical infrastructure (e.g., autonomous industrial robots, transport systems).
 - Worker management tools (e.g., biometric monitoring, automated scheduling).
 - AI applications in migration, law enforcement, education, or access to essential services (e.g., credit scoring).

Obligations for providers and deployers include comprehensive risk assessment, robust data quality standards, traceability (logging), detailed technical documentation, human oversight, transparency measures and ensuring accuracy, cybersecurity and robustness.

3. **Limited risk:** Systems requiring transparency obligations (e.g., chatbots, generative AI) to ensure users know they are interacting with AI or consuming AI-generated content. Providers must also label deepfakes and other AI-generated content where appropriate.
4. **Minimal or no risk:** Most AI applications (e.g., spam filters, AI-enhanced video games) fall into this category and are not subject to additional regulation under the AI Act.

Furthermore, the AI Act introduces specific rules for General-Purpose AI Models (GPAI Models), also known as foundation models (Schneider, 2024), reflecting growing regulatory attention to their broad applicability and systemic impact. These models, typically trained on

¹⁷ AI Index Report 2025. https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf

¹⁸ European Parliament & Council of the European Union. (2024). *Artificial Intelligence Act, Regulation (EU) 2024/1689*. *Official Journal of the European Union*, L 259, 1-157.



large datasets through self-supervised learning, can be deployed across a wide range of applications, including chatbots, recommendation systems, industrial robotics and medical diagnostics.

The Act distinguishes between:

- Standard GPAI Models, which are subject to baseline obligations (e.g. documentation, transparency, copyright compliance) and
- GPAI Models with systemic risk, which must meet enhanced requirements, such as conducting model evaluations, risk mitigation and reporting on energy use and safety measures.

These obligations apply regardless of the provider's location, provided the model is placed on or used in the EU market.

Industrial AI systems are often classified as high-risk under the AI Act, primarily due to their safety-critical functions or impact on fundamental rights. According to the AI Act, high-risk AI systems include those that:

- Are used as safety components of products (e.g., autonomous industrial robots, AI-controlled assembly lines);
- Fall under harmonised product legislation (e.g., Machinery Regulation, Medical Devices Regulation);
- Are used for worker management, such as AI tools for productivity tracking, biometric monitoring, or automated scheduling;
- Are deployed in critical infrastructure sectors like energy, transportation, or logistics.

High-risk classification triggers a cascade of obligations for providers and deployers, aimed at ensuring safety, accountability and legal compliance throughout the lifecycle of the AI system.

The AI Act imposes a number of ex ante and post-market obligations for high-risk AI systems. These are relevant both for AI providers and deployers.

a. **Risk management (Art. 9)**

A comprehensive risk management system must be established and maintained. This includes identification, analysis and mitigation of foreseeable risks, such as physical harm from malfunctioning machinery or legal risks from unfair treatment of workers by AI scheduling tools.

b. **Data governance (Art. 10)**

Training, validation and testing datasets must be relevant, representative, free of errors and it must consider potential biases. This is particularly important in industrial AI applications involving human-machine interaction, where biased data could lead to unsafe or discriminatory outcomes.

c. **Technical documentation (Art. 11)**

Providers must maintain documentation describing system architecture, intended purpose, training methodologies and post-deployment controls. This documentation supports regulatory audits and ensures traceability.

d. **Logging and record-keeping (Art. 12)**



High-risk AI must log operational events in a way that allows traceability and accountability. In industrial AI, this may involve recording decision paths in automated quality control systems or documenting the actions of collaborative robots.

e. Transparency and instructions for use (Art. 13)

Industrial deployers must be informed of the AI system's capabilities and limitations. For example, if an AI tool is used to detect defects in manufacturing, users must understand under what conditions it might underperform.

f. Human oversight (Art. 14)

High-risk AI must be subject to effective human intervention mechanisms, allowing the system to be overridden if necessary. In automated factories, this implies that humans must be able to interrupt the operation of autonomous machines to prevent harm or system failure.

g. Robustness, accuracy and cybersecurity (Art. 15)

Systems must operate reliably and be protected from cybersecurity threats. For AI deployed in industrial IoT settings, this involves safeguards against system hacking, adversarial attacks or data manipulation.

h. Conformity assessment and CE marking (Arts. 16-23, 49)

Before deployment, high-risk AI must undergo a conformity assessment, either through internal control (self-assessment) or via a notified body. Following successful assessment, the system receives the CE marking, indicating its compliance with the AI Act.

The AI Act represents a landmark step in EU digital regulation, aiming to align industrial innovation with safety, legal certainty and accountability. For industrial users of AI, the Act mandates a structured compliance framework that encompasses technical robustness, transparency, documentation and human oversight. As AI systems become increasingly integrated into industrial processes, companies must adapt to a dynamic regulatory environment where legal compliance functions as a core component of responsible and scalable digital deployment.

3.4 Ethical considerations and challenges on using AI in industry

The use of AI technology in industry raises significant ethical considerations and challenges, particularly as AI systems become increasingly autonomous and influential in decision-making processes. The main ethical aspects involved in the industrial use of AI include issues such as accountability, transparency, bias mitigation, data protection and the potential impact on the environment and global inequality.

3.4.1 Transparency and explainability

The requirement for transparency and explainability in AI systems is central to building trust and ensuring accountability, particularly in industrial settings where AI systems influence decisions with economic, safety and human implications. In the EU legal framework, the AI Act mandates that high-risk AI systems must be sufficiently transparent to enable users to interpret and appropriately use their outputs (Article 13 AI Act). This includes providing information on the system's capabilities, limitations and how it should be used. The lack of explainability, especially in complex "black-box" models like deep learning, can undermine both legal compliance and ethical acceptability (Nastoska et al., 2025).

Furthermore, GDPR (Art. 22) grants individuals the right to not be subjected to solely automated decisions with significant effects unless there is meaningful human intervention or explanation. This is crucial in industrial contexts such as automated hiring or performance assessments, where AI decisions have direct economic or safety impacts.

The right to good administration under the EU Charter (Art. 41) also requires transparency and reason-giving for decisions, reinforcing ethical imperatives for fairness and accountability.

From an ethical perspective, explainability supports human autonomy, fairness and accountability. It ensures that operators, managers and affected individuals can understand and contest AI decisions, fostering both operational safety and human dignity. To operationalize transparency, the EU encourages the use of model documentation (model cards), user instructions and auditable logs, as specified in Annex IV of the AI Act.

3.4.2 Fairness and non-discrimination

Ensuring fairness and preventing discrimination are fundamental ethical imperatives in the deployment of AI systems in industry, especially when such systems are used in employment, resource allocation or supply chain management. Industrial AI tools, whether for recruitment, worker evaluation or operational decision-making, can replicate or amplify existing societal biases if not properly designed and audited.

The AI Act explicitly requires that high-risk AI systems be developed and used in a manner that minimizes risks of bias and discrimination (Article 10). Providers of such systems must implement data governance and data quality controls, ensuring that training, validation and testing data are representative, relevant and free from discriminatory bias. This is reinforced in Annex IV, which includes documentation requirements about how systems are trained and tested for fairness.

Under EU anti-discrimination law (e.g., Directive 2000/78/EC, the Employment Equality Directive), employers are prohibited from discriminatory practices on the basis of age, gender, race, disability, religion, or sexual orientation. Where AI tools are used in employment-related decisions, such uses must comply with these principles. Importantly, liability for discrimination may still arise even if the bias is embedded in the algorithm, not in the human user.

The GDPR also supports fairness through Article 5(1)(a), which requires data to be processed lawfully, fairly and in a transparent manner.

Ethically, fairness entails designing AI systems that do not unduly advantage or disadvantage any individual or group. It also implies the need for ongoing auditing, inclusive stakeholder engagement and mechanisms for redress. Without such safeguards, AI in industry risks reinforcing existing power imbalances, particularly between employers and workers.

3.4.3 Privacy and data protection

Industrial AI often processes large amounts of personal data, including sensitive information from employee monitoring, wearables, or customer interactions. This processing must meet the strict conditions set by the GDPR (Arts. 6, 9), including lawful purpose, data minimization and purpose limitation (Article 5(1)(c) and Article 5(1)(b) respectively).

The AI Act complements GDPR by mandating privacy-by-design and by-default measures (AI Act, Art. 9), such as encryption, anonymization and strict access controls. These technical and

organizational safeguards aim to protect data subjects from intrusive or disproportionate surveillance.

Additionally, individuals retain rights to access, rectify and object to automated processing that affects them (GDPR Arts. 15, 16, 22). For example, an employee subject to an AI-driven productivity review must have the right to understand and challenge the evaluation.

Ethically, data processing in industrial AI must respect informational self-determination, especially in contexts with unequal power dynamics like employer-employee relations, requiring clear transparency and freely given consent or alternative lawful bases.

3.4.4 Human oversight and autonomy

Maintaining meaningful human control over AI decisions is a cornerstone of both EU law and AI ethics. The AI Act requires high-risk AI systems to have mechanisms enabling human intervention or override (Art. 14), whether through real-time control (“human-in-the-loop”), monitoring (“human-on-the-loop”), both referring to active human involvement during AI operation (Fink, 2025), or ultimate decision authority (“human-in-command”).

Moreover, Recital 47 of the AI Act emphasizes that human oversight is a safeguard to preserve fundamental rights, including dignity and freedom of choice. This is especially important in employment scenarios, where algorithmic tools may influence hiring, task allocation, or productivity monitoring. These decisions can significantly affect a person’s livelihood and must not be fully delegated to autonomous systems.

Complementary safeguards exist in the GDPR, particularly Article 22, which grants individuals the right not to be subject to solely automated decision-making with legal or similarly significant effects unless suitable safeguards, such as the right to human intervention are in place.

Effective oversight also requires organizational governance including clear accountability, operator training and procedural safeguards to integrate AI responsibly into workflows.

3.4.5 Responsibility and accountability

As AI becomes deeply embedded in industrial operations, clarifying who is responsible for AI-driven outcomes is critical.

The AI Act delineates responsibilities between AI system providers (developers) and users (deployers), with providers obligated to ensure conformity assessments, risk management and documentation, while users must operate AI systems as intended and report malfunctions (Arts. 16-29). This division of roles supports the principle of traceability, a core requirement under Article 12 of the AI Act, which mandates that systems be developed in a manner that allows decisions or outcomes to be tracked and explained.

From an ethical perspective, accountability in AI involves more than just legal liability. It includes moral responsibility to design systems that prevent harm, ensure fairness and allow redress. Without clear lines of responsibility, industrial actors may be tempted to “blame the algorithm” when failures occur (Budnik, 2025). To counter this, companies must adopt AI governance frameworks that define responsibilities internally (e.g., compliance officers, ethics boards) and externally (e.g., supplier obligations).

3.4.6 Environmental impact and global inequality

While AI is often promoted as a tool for industrial optimization and sustainability (e.g. predictive maintenance, resource-efficient logistics), its use also raises serious concerns about environmental sustainability and global fairness. Industrial AI systems require substantial computational resources, infrastructure and energy, especially in training and operating large-scale models or real-time analytics platforms.

The AI Act does not yet impose binding environmental requirements but acknowledges the importance of sustainability (Recitals 6 and 47), urging energy-efficient design. EU policies such as the European Green Deal¹⁹ and the Corporate Sustainability Reporting Directive²⁰ require companies to report on environmental impact, including AI's carbon footprint and electronic waste.

Ethical concerns extend to global inequalities since AI supply chains depend heavily on rare earth minerals often sourced under exploitative conditions in low-income countries. The concentration of AI technology in wealthy regions risks deepening global digital divides and economic disparities.

Industrial AI development should therefore integrate life cycle assessments, minimize resource consumption, promote recycling and ensure supply chains are socially and environmentally responsible. Frameworks like UNESCO's Recommendation on the Ethics of AI support these principles by advocating for equitable, sustainable and inclusive AI governance²¹.

¹⁹ <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2019:640:FIN>

²⁰ <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022L2464>

²¹ <https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence>

4 Impact and challenges of AI deployments

4.1 Economic impacts of AI deployments

Whilst during previous eras where AI flourished during the 20th century, such as from the 1970s to the 1990s, it was mostly constrained in academic research settings funded by governments, in the current boom AI has been propelled mostly by investors in the private sector (Haigh, 2025). This turn towards AI products geared for the market has been boosted by the success of LLMs which entered the public sphere on the 30th November 2022, with the release of OpenAI's ChatGPT, a generative AI chatbot. The success of ChatGPT and subsequent increasingly more sophisticated LLMs has contributed to the sky-rocketing of the value of the whole AI sector. Currently, the value of the AI market size stands at USD 757,58 billion²². According to the same analysis, it is projected to reach a value of USD 3.680,47 billion by 2034. According to the UK government's own statistics, the AI sector is growing 30 times faster than the rest of the economy²³ whilst in the EU, AI companies represented 20% of the whole of venture capital funding in 2024 (Dillet, 2025).

AI is predicted to fuel economic growth, with the most enthusiastic analysts branding it as a revolutionary tool that will transform technology and society on a par with the industrial revolution in the 19th century and the computing revolution in the 20th century. An influential 2023 report by Goldman Sachs Research (Briggs & Kodnani, 2023) claimed that advances in the AI sector would drive global GDP upwards by up to 7% over a ten year period and lift productivity growth by approximately 1,5% over the same period. According to the authors, this effect will occur because of the macroeconomic effects of generative AI's increasing ability in generating content which is indistinguishable from human-generated content, as well as the breaking down of the barriers in communication between humans and machines.

On the other hand, however, 2024 Nobel Laureate for Economics sciences, Prof. Daron Acemoglu, claimed in a 2025 paper (Acemoglu, 2025) that the effects of AI on the global economy would be non-trivial yet relatively modest, with a prediction of an increase of 0,53% in total factor productivity (TFP). He argues that the exaggeration in the various triumphant claims and predictions is due to the successes of AI in what he labels 'easy-to-learn' tasks, whereas success in the future will have to come from 'hard-to-learn' tasks in which the decision-making will be more complex due to contextual factors and where the indicators of successful performance will be much more difficult to discern and measure.

Another significant economic (and societal) impact of AI is the impact on the employment market. According to the Goldman and Sachs report, approximately two thirds out of 900 occupations listed in US databases have been affected by the advent of automation brought by AI. In terms of jobs, it is argued that 300 million jobs could be affected by AI in the next decade.

However, this does not mean that workers will be left with no choice in the labour market, nor is this the first time that such a displacement due to technology has happened. The labour market is fluid and most occupations currently in existence did not exist prior to the 1940s. It

²² <https://www.precedenceresearch.com/artificial-intelligence-market>

²³ <https://www.gov.uk/government/news/pm-launches-national-skills-drive-to-unlock-opportunities-for-young-people-in-tech>

is thus claimed that new occupations will emerge, which will enable humans to use new skills and abilities. Finally, there are sectors such as healthcare in which AI may not be able to displace humans due to unique (or at least, not replicated by AI to a significant degree yet) qualities that humans possess, such as emotions.

Another theoretical and indirect economic impact of AI is that it may be claimed that it has significantly altered the workings of the dominant economic system. According to some commentators, for example Varoufakis (2023) and Swyngedouw (2025), in roughly the last two decades, the most dominant economic system is no longer the late-twentieth century capitalist one but a new system variably called technofeudalism, capitalist feudalism or new feudalism. This system is characterised by the fact that a few individuals and the companies they own are valued more than most nation-states' GDP (for example Google, Alibaba, Amazon, Meta, Apple, Microsoft) and they mostly derive their wealth from rent extraction and the commodification of services extracted by the users of their platforms themselves (Swyngedouw, 2025). Whilst this change was not triggered by AI itself, AI exacerbates inequality as these big private technological corporations are the main investors with the needed resources to maintain and further the effects of the AI revolution.

4.2 Societal and ethical impacts of AI deployments

The key societal and ethical challenges which are at the heart of ALFIE are bias and fairness. Though it is true that it is impossible to avoid bias, AI can perpetuate or amplify biases present in training data, leading to discriminatory outcomes in areas like hiring, lending, policing, privacy and surveillance. AI-driven surveillance technologies, facial recognition and data mining raise privacy and civil liberty concerns. According to a 2022 report commissioned by the OECD, 97 out of 179 countries use AI and big data surveillance tools (Feldstein, 2022). The report cites 64 countries using smart city / safe city platforms, 78 countries using public facial recognition systems, 69 countries using AI and big data tools for smart policing and 38 countries using social media surveillance tools. The report, however, does not distinguish between legitimate and illegitimate uses of such tools.

Another common concern is that of AI eliciting misinformation and deepfakes, since increasingly AI can generate convincing fake content, leading to challenges in identifying truth and spreading misinformation. A demonstration of the application of deepfakes to spread misinformation occurred in 2024 when images allegedly taken after the storming of an infamous Syrian prison were later revealed to have been AI-generated²⁴.

Finally, the economic problem of job displacement, mentioned above, also has societal implications. The threat of massive job displacement, particularly in sectors involving routine cognitive or physical tasks (Rawashdeh, 2025), may affect the welfare of large swathes of vulnerable populations through marginalisation and mass unemployment. The actual threat is not AI, but the urgent need for coordinated action among stakeholders (regulators, policy makers, educators and governance). Policy recommendations should focus on talent retention, investment in upskilling programs and the establishment of support mechanisms for those adversely affected by AI adoption, since there is a global shortage of professionals who can develop, deploy and manage AI systems responsibly.

²⁴<https://www.dw.com/en/fact-check-fakes-on-the-rise-after-rebels-open-saydnaya-prison-in-syria/a-71030605>

A final class of related societal problems is that of security risks. AI systems can be attacked or manipulated (e.g., via data poisoning, model inversion), posing cybersecurity risks. Given the reliance of institutions such as banking institutions, scientific institutions and state and supra-state institutions, these problems should be always kept in mind and never be underestimated.

4.3 Limitations and challenges of current deployments

Current AI deployments offer powerful capabilities, but they face several limitations and challenges across technical, ethical and operational domains. Regarding technical limitations there is a lack of General Intelligence (Artificial General Intelligence or AGI). Most AI systems are narrow and specialized. They excel at specific tasks (e.g., image recognition, language translation) but fail at broader reasoning or adapting across diverse tasks (Mumuni & Mumuni, 2025). So multitasking is one of the limitations, along with nuance, ambiguity, or deeper contextual understanding. The key challenge in multitask learning by AI is that of combining learning signals from different tasks into a single model. Nuance and ambiguity refer to human linguistic phenomena which AI models have failed to master so far, whilst deeper contextual understanding refers to the cognitive and linguistic abilities required to understand linguistic, cultural and situational context within a conversation (Qamar et al., 2025). Whilst humans may draw on life experience, social cues and nuanced knowledge to understand phenomena like irony and emotional language, AI models lack these tools and only rely on surface correlations in linguistic data. The difficulties described above render human-computer interaction still not a seamless phenomenon, whilst at the same time they affect trust in AI decision-making.

However, the issue of AI has stirred quite a number of high-profile reactions and predictions both from scholars and from big tech AI firm owners and technology enthusiasts in the public arena in the last few years - these reactions vary on the technological optimism-pessimism spectrum, but also bring to the fore more primordial fears and attitudes regarding the relationship between humans and AI. Two particularly pointed high-profile examples include Sam Altman on the optimist side²⁵ and Sir Demis Hassabis on the pessimist side²⁶.

Another issue is that AI can generate plausible text or outputs that are actually incorrect or misleading. This is also directly linked to data dependency since AI performance is highly dependent on the quality, quantity and diversity of data. In simpler terms, AI models are more or less as good as the data they are trained with. Instances where AI generates intelligible yet factually incorrect answers are labelled as hallucinations and these are occurring even in more recent, theoretically superior models (Z. Zhang, 2025, Bang et al., 2025).

The related technical notions of explainability and interpretability are also issues to be considered since many AI models, especially deep learning models, are technological "black boxes". Technological black boxes, as termed by the sociologist of scientific knowledge Bruno Latour (1987) refer mostly to technological devices (but also to other things such as scientific processes etc.) which receive an input and give out an output, without it not being clear what happens in between these two phases. Given the use of LLMs by lay people in their daily lives, as well as in critical specialised domains such as healthcare, education, the legal system

²⁵ <https://firstmovers.ai/agi-2025/>

²⁶ <https://www.windowcentral.com/artificial-intelligence/google-deepmind-ceo-dismisses-claims-of-phd-level-ai-as-nonsense>



and finance, the difficulty to interpret or audit their outputs or decisions becomes a major concern. Explainability is a practical and user-focused notion linked to the user's need and right to understand decisions that affect them, as well as to the notion of algorithmic transparency. Satisfying the need for explanation allows for the building of trust, as well as to the contestation of the provided explanations. Interpretability is a stronger notion, as it is structural and is linked to ethical values such as transparency, safety, fairness and accountability, as well as the cognitive value of genuine understanding. Significant research steps have been taken towards AI interpretability, with a recent review citing the progression from behavioural (black box), to attributional, to concept-based and finally to the mechanistic paradigm in interpretability (Bereska & Gavves, 2024).

Robustness is another limitation of AI. It is defined as the ability of a system to cope with erroneous input and to cope with errors during command execution. AI systems often lack robustness to adversarial inputs or unexpected data. Small changes in input can cause them to fail unpredictably. This is a severe limitation of these systems as it deeply affects trust in their decisions and outputs. In other words, the robustness of AI affects its reliability. Reliability in turn is crucial for applications in domains such as the law, medicine and weather forecast and climate change modelling.

Another challenge for AI models is their high computational costs, especially during training. The high computational costs associated with training and deploying cutting edge AI models means that they require a massive input of computational power and energy, something which raises climate change and sustainability concerns.

AI models also face the challenge of the financial cost involved in their creation, mostly during their training phase. This cost was estimated in data in Forbes magazine (Buchholz, 2024) in the ballpark of 100-200 million dollars for the most advanced model (Google's Gemini 1) in August 2024 and it is projected to escalate rapidly for future models as they become more sophisticated. A further reason for the sky-rocketing of prices is the arms race between tech companies and the demand in various cutting edge high-stakes domains such as military uses or scientific uses. An example of such nefarious use of AI is the recent war in Gaza, where the Israeli army has purportedly used two AI tools to gather intelligence and identify targets (Davies et al., 2023).

Finally, a lack of diversity and richness in the AI model training data, or even downright biased data leads to poor or biased outcomes. Identifying the bias is a necessary first step. However sometimes further steps needed to correct the bias, such as finding unbiased datasets or uncovering the underlying assumptions behind it and then in turn retraining the model in such a way as to eliminate it is not such an easy endeavour. Current AI deployments face bias at multiple levels — data, algorithm, human feedback and deployment — and these biases are difficult to detect, mitigate and regulate. As AI adoption grows, still beyond more explainable models, stronger regulations and ethical frameworks co-developed with affected communities the main concern seems to point to the lack of standardisation of frameworks for measuring and mitigating bias across industries and jurisdictions. After the entry into force of the AI Act in August 2024, an open question identified by the EU is its interplay with the GDPR. The AI Act aims to promote human-centric, trustworthy and sustainable AI, while respecting individuals' fundamental rights and freedoms, including their right to the protection of personal data. One of the AI Act's main objectives is to mitigate discrimination and bias in the development, deployment and use of 'high-risk AI systems'. To achieve this, the act allows 'special categories of personal data' to



be processed, based on a set of conditions (e.g. privacy-preserving measures) designed to identify and to avoid discrimination that might occur when using such new technology. The GDPR, however, seems more restrictive in that respect. The legal uncertainty this creates might need to be addressed through legislative reform or further guidance. Hence inconsistent policies on fairness, transparency and accountability are at the heart of the main concerns leading to difficulty balancing fairness with accuracy, privacy and performance.



5 Explaining AI models to mitigate bias and ethical concerns

The adoption of AI has raised fundamental concerns regarding fairness, transparency and ethical accountability. A good example is the results from evaluating gender bias in large language models in long-term care (Rickman 2025). Central to these concerns is the lack of explainability in many state-of-the-art models. Deep learning systems, for instance, can generate highly accurate predictions, yet often do so through opaque processes that users and even developers struggle to interpret. This “black box” nature of AI limits the ability to detect biased behaviour, contest decisions, or ensure ethical standards, especially when systems are deployed in sensitive, high-stakes environments.

Bias in AI can emerge through skewed training data, flawed model assumptions, or unequal feature representations (Hasanzadeh, et al., 2025). Without interpretability, such bias can remain hidden, leading to outcomes that reinforce discrimination and social inequity. Consider algorithmic hiring systems or predictive policing tools, both have faced criticism for perpetuating historical injustices due to unexamined data patterns. The challenge becomes even more pressing in applications like autonomous driving or medical diagnostics, where misjudgements can pose physical risks. Thus, explainability is not simply a technical feature but a necessary ethical function that enables accountability, human oversight and compliance with regulatory and societal expectations.

Several recent studies reinforce this point. A 2023 review by Notovich, et al. (2023) emphasized that explainability plays a vital role in demystifying complex AI outputs and restoring trust among stakeholders. They outline how explainability should be seen not as an optional add-on but as an integrated design principle, particularly in contexts that demand transparency for safety or fairness. Similarly, the 2024 report *Mapping the Landscape of Ethical Considerations in Explainable AI* (Nannini, et al., 2024) stresses the social contract around AI systems and argues for embedding explainability into legal and ethical frameworks from the start. Without such measures, AI runs the risk of becoming an unaccountable and unjust decision-making tool.

While efforts to regulate AI systems continue to mature, explainability has become a pivotal mechanism to enable auditability, contestability and human-in-the-loop decision-making. It allows affected individuals and regulators to understand why an outcome was produced and whether the model's behaviour aligns with both legal standards and human values. This is especially important as AI becomes more autonomous and influential in shaping real-world decisions.

5.1 Current technologies for explainable AI

Over the past few years, a variety of technical approaches have emerged to provide interpretability for both simple and complex AI systems. Broadly speaking, these methods can be categorized into two types: intrinsic (or interpretable-by-design) and post hoc explainability techniques.

Intrinsic methods involve models that are transparent in their structure and logic. For example, decision trees, logistic regression and rule-based systems are inherently interpretable, as each decision pathway or coefficient can be traced and explained. These models offer clarity in

domains where explainability is prioritized over raw predictive power. However, they may fall short when applied to tasks requiring complex, nonlinear representations, such as image recognition or language understanding.

In contrast, post hoc explainability techniques are used to interpret black-box models after they have been trained. Among the most widely used are SHAP²⁷ and LIME²⁸ (Local Interpretable Model-Agnostic Explanations). SHAP applies cooperative game theory to attribute contributions of each input feature to a model's prediction in a consistent and theoretically grounded manner. LIME, on the other hand, builds simple local surrogate models to approximate how a complex model behaves around a specific prediction. These tools have found wide adoption across industry and academia, particularly for debugging, compliance audits and building user trust.

In computer vision and biomedical applications, visual attribution techniques like Grad-CAM are frequently employed. These highlight the regions of an image that a convolutional neural network has focused on when making a classification decision. For example, in clinical imaging, such visual maps not only aid clinicians in verifying results but also serve as a form of ethical reassurance that the AI is attending to medically relevant features. Mathew, et al., (2025) argue that such visual explanations have improved user confidence in AI-assisted diagnostic platforms by offering traceable evidence of model reasoning.

More recently, advances in mechanistic interpretability have attempted to go beyond feature attribution and toward understanding the actual internal logic of deep neural networks. These efforts involve dissecting network components to reveal functional substructures, such as neurons responsible for specific tasks or reasoning pathways. While still a developing field, this direction holds promise for rendering even the most complex AI systems intelligible to humans.

In addition to these methods, iterative bias detection and correction frameworks have gained momentum. For instance, researchers have proposed life-cycle models where explanations are used not just to interpret outputs but to reveal and revise biased training data and flawed model assumptions. The MICCAI 2023 conference²⁹ highlighted one such “reveal-to-revise” framework where explainability becomes a central component of model refinement over time, with flagship studies as the one of Pahde et al. (2023). Rather than offering explanations only at the end, the system uses them throughout development to identify and correct problematic behaviours iteratively.

However, it is crucial to recognize that explainability alone is not sufficient to ensure fairness. As Deck, et al. (2023) emphasized in their critical survey, many explainability tools risk giving a false sense of security if they are not tightly integrated with fairness metrics and stakeholder-centered evaluation. The effectiveness of XAI depends on how explanations are framed, who receives them and whether they are actionable. In contexts where power imbalances or technical illiteracy exist, raw technical explanations may be insufficient or even misleading.

Ultimately, explainability technologies are most impactful when deployed as part of a broader ethical and governance framework. Their purpose is not merely to open the black box but to

²⁷ https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html

²⁸ <https://c3.ai/glossary/data-science/lime-local-interpretable-model-agnostic-explanations/>

²⁹ <https://conferences.miccai.org/2023/en/>



make AI systems contestable, improvable and aligned with human values. As regulations such as the EU AI Act evolve and public scrutiny increases, integrating robust explainability into the AI life cycle will be vital for ensuring responsible innovation.

5.2 The role of explainability in AI

Explainability plays a foundational role in ensuring the transparency, accountability and ethical acceptability of AI systems. As AI becomes increasingly embedded in critical decision-making domains, such as healthcare, criminal justice, finance and autonomous vehicles, there is a growing need for models that not only perform well but can also be understood by humans. Explainability enables stakeholders to trace how and why a model arrives at a particular output, thereby allowing for informed oversight, trust and correction when necessary. It bridges the gap between complex algorithmic reasoning and human comprehension, particularly when high-stakes outcomes are involved.

One of the primary motivations for XAI is to identify and mitigate potential biases that may be embedded in training data or model behaviour. As Hasanzadeh, et al. (2025) emphasised, bias in AI often operates silently, reinforcing historical inequalities unless explicitly exposed and addressed through interpretable frameworks. Furthermore, explainability is essential for aligning AI decisions with legal and ethical standards, especially in jurisdictions that demand the right to explanation under data protection regulations. It also facilitates human-AI collaboration by allowing users to trust model outputs based on reasoned understanding rather than blind faith.

Recent scholarship underscores the strategic function of explainability throughout the AI lifecycle, from data selection and model design to evaluation and post-deployment monitoring. For example, Nannini et al. (2024) highlight that explainability should not be viewed solely as a technical feature, but as a normative requirement that shapes responsible AI governance. By enabling developers, regulators and users to interrogate model logic, explainability supports fairness, improves safety and contributes to AI systems that are both effective and socially aligned.

5.3 Ethical challenges and bias in explainable AI

XAI promises increased transparency and trust in algorithmic decision-making, yet it also introduces a set of ethical challenges that must be carefully managed. One prominent issue is cultural and social bias, where explanations generated by XAI systems often fail to consider diverse user backgrounds. A recent semi-systematic review found that most XAI research focuses on Western populations and frequently neglects underrepresented communities, resulting in culturally insensitive or incomplete explanations (Thalpage, et al., 2025). Without inclusive design practices, XAI may inadvertently replicate the same inequities it aims to mitigate.

Additionally, there is a risk of moral outsourcing, where the opaqueness of technical explanations allows developers to offload ethical responsibility onto algorithmic systems, thereby evading accountability. This effect is exacerbated when individuals or organizations interpret XAI explanations as moral absolution rather than starting points for human-led ethical reflection.

A further concern is adversarial manipulation: detailed explanations such as feature importance scores or saliency maps, may offer strategic insights that can be exploited. For instance, knowledge of which demographic features affect loan or hiring decisions can be used to game the system, thereby deepening existing inequalities (Bhatt, et al., 2020). Achieving a balance between transparency and security remains a persistent challenge.

XAI explanations may also foster false confidence among users. If explanations are too simplistic or superficially convincing, users may trust the model blindly despite underlying biases or inaccuracies. This *trust without understanding* dynamics undermines the very accountability XAI is intended to promote.

Domain-specific constraints also complicate ethical deployment. In areas like healthcare and criminal justice, explanations must not only be accurate but also comprehensible to practitioners and patients. Studies indicate that XAI methods tailored to global models may not translate effectively in these high-stakes settings without careful domain-aware adaptation (Gohel, et al., 2021).

Given these challenges, ethical XAI must evolve beyond transparency alone. It should integrate participatory design, where diverse user groups contribute to explanation development, alongside governance mechanisms that ensure ongoing ethical monitoring. Additionally, techniques like explanation redaction and adaptive explanation levels can help reduce gaming risks while maintaining interpretability.

In essence, the ethical promise of XAI lies not just in generating explanations, but in ensuring these explanations are culturally sensitive, secure, responsibly used and subject to continuous human oversight.

5.4 Ethical consideration and gap analysis

Ethical considerations in XAI are increasingly recognized as essential, yet current implementations often fall short of embedding ethical rigor into system design. A recent critical review found that while many XAI papers mention ethics, discussions are often superficial and lack engagement with established ethical theories such as deontology or virtue ethics (Nannini, et al., 2024). This illustrates a troubling gap: transparency is celebrated in principle, but rarely contextualized with real-world moral complexity or stakeholder perspectives.

An especially concerning gap exists between ethical principles and their operationalization. As observed in global analyses of AI ethics, high-level ethical frameworks, though widely endorsed, frequently fail to translate into practical measures in deployed systems (Suresh & Guttag, 2021). In XAI specifically, this translation deficit means explanations may be technically coherent without offering meaningful ethical insight or enabling accountability.

Another underexplored area is the moral value of explanations themselves. Researchers argue that XAI should not just facilitate understanding, but also support reciprocal moral obligations, empowering users to challenge decisions and hold systems accountable. Without these moral dimensions, explanations risk becoming perfunctory, satisfying forms without substance.

Furthermore, critical analysis reveals that many XAI systems present oversimplified fairness claims without clarifying which fairness criteria are affected or who benefits, overlooking the multi-faceted nature of justice (Deck, et al., 2023). This ambiguity can have unwanted practical



repercussions: decisions framed as “fair” may still perpetuate inequality when fairness is poorly defined or contextualized.

Finally, while XAI is frequently proposed as a tool for ethical governance, there remains little structured guidance on how to embed it across the AI lifecycle, from data collection and model design to deployment and monitoring. This absence of lifecycle integration creates a misalignment between ethical aspirations and operational practices (Mittlestadt, et al., 2019).

5.5 Future directions for resolving challenges and ethical problems

As AI continues to influence increasingly sensitive domains such as healthcare, education, law enforcement and public policy, the limitations of current ethical safeguards and explainability methods have become more apparent. Moving forward, the focus must shift from reactive transparency toward proactive, socially responsive design. Future developments in AI ethics and explainability must address not only how we interpret model decisions, but also who these interpretations are for and under what conditions they meaningfully support accountability and trust.

One key direction involves developing context-sensitive explainability. Rather than producing uniform outputs, future XAI systems must generate explanations that reflect the expertise, expectations and cultural context of different user groups. For instance, medical professionals require different types of explanations than patients or regulatory bodies, even when discussing the same model outcome. Designing explanation systems that adapt to these diverse needs can help reduce misinterpretation and bridge communication gaps between AI systems and human stakeholders.

Security-aware explainability will also be crucial. As current methods risk being exploited, intentionally or not, by those seeking to manipulate decisions or obscure accountability, future models must include safeguards against adversarial misuse of explanations. Balancing transparency with protection from gaming will require nuanced strategies such as conditional disclosure, explainability thresholds, or layered explanatory content that adapts based on usage risk.

In parallel, governance frameworks will need to evolve. Future AI systems should be developed with built-in auditability, documentation and continuous monitoring processes that extend well beyond initial deployment. These mechanisms must be supported by regulatory clarity and organizational structures that assign responsibility for ethical oversight throughout the AI lifecycle. Ethical AI will not emerge solely from better code or algorithms, it will require sustained cooperation between developers, legal experts, policymakers and affected communities.

Ultimately, the future of resolving ethical and bias-related challenges in AI lies in designing systems that not only reveal how they work but align with the values and expectations of the societies they serve.

6 Use of speech-to-text and NLP in AI and its ethical considerations

6.1 State-of-the-art NLP models

Recent advancements in Natural Language Processing (NLP) are predominantly driven by Large Language Models (LLMs) and their smaller, efficient counterparts, Small Language Models (SLMs) (Q. Zhang et al., 2025). These models have revolutionized tasks such as writing, coding and text-based activities, with notable examples including OpenAI's "o-series" (o1, o3, o4-mini), GPT-4o and open-source models like Nemotron-4-340B or Llama 3. LLMs achieve remarkable proficiency through extensive training on diverse datasets, including publicly available internet data, partnered third-party information and multilingual content spanning over many languages (Nvidia, Adler et al., 2024). The NLP landscape is divided between open-source and closed-source models. Open-source models foster community engagement, customization and transparency. In contrast, closed-source models like GPT-4 and OpenAI's o-series offer robust performance but limited accessibility to their internal mechanisms.

Open-source large language models such as Llama-2-70B (Touvron et al., 2023) release not only the final checkpoint but also the training code, tokenizer and hyper-parameter recipes, which permits researchers to reproduce the pre-training corpus filtering pipeline, verify the positional embedding implementation and even fine-tune on domain-specific corpora, whereas closed-source models such as GPT-4 (OpenAI, 2023) expose only black-box token-level APIs, thereby preventing the community from auditing whether the reported architecture or the context window are actually instantiated as claimed.

Also security and compliance audits are fundamentally asymmetric: with open checkpoints one can run offline red-teaming, inspect every weight for watermarking or back-door triggers and ship an air-gapped copy to a classified environment (Vassilev et al., 2025), whereas closed-source systems require sending potentially sensitive prompts to a remote fleet whose firmware, logging policy and side-channel defences are difficult to audit or even unauditable black boxes (Nasr et al., 2023).

Context size, a critical factor in model coherence and understanding, continues to expand (Nvidia, Adler et al., 2024). Larger context windows enable better comprehension but increase computational demands. Techniques such as grouped query attention (GQA) (Nvidia, Adler et al., 2024), efficient attention mechanisms and sparse Feed-Forward Networks (FFNs) (Z. Zhang et al., 2025) are being developed to mitigate these challenges of increase for computational power demands. SLMs, created through model compression techniques (quantization, pruning and knowledge distillation), offer solutions for efficient deployment on resource-constrained devices.

Multilingual capabilities are inherent in many LLMs due to diverse training data. However, performance variability across languages persists, with English often benefiting from more extensive training datasets. Addressing this disparity is essential for truly inclusive NLP solutions. Models like mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) have made strides in multilingual representation, but under-resourced languages still lag behind. Techniques such as cross-lingual transfer learning and data augmentation (e.g., back-translation) can help bridge this gap. The gap between languages (or, more specifically, the

amount of data available to develop the models) does not only affect the performance of the models, but also operating costs. For instance, while processing a sentence in English may require 50 tokens, processing its counterpart translated on a limited-resources language like Slovak or Greek may require 1.5 to 2 times as many tokens, which directly translates into increased computational complexity and costs.

Multimodality, the integration of text with other data types (e.g., images, audio), is an emerging focus area, exemplified by models like Gemini. The exploration of multimodal capabilities is increasingly crucial for comprehensive language understanding. For instance, models such as CLIP (Radford et al., 2021), GPT-4o and Gemini-1.5-Pro have shown promising results in vision-language tasks, demonstrating the potential of multimodal training. Multimodal models can enhance tasks such as visual question answering and cross-modal retrieval.

LLMs are increasingly finding applications within Automated Machine Learning (AutoML) systems, demonstrating a significant shift in how these automated processes function (Tornede et al., 2024). Research indicates LLMs can be leveraged for AutoML itself, acting as components to streamline and enhance the automation of machine learning pipelines (Tsai et al., 2023).

One key application is using LLMs to guide hyperparameter optimization (HPO), a crucial step in building effective ML models (S. Liu, Gao and Li, 2024). These models can assist in suggesting optimal configurations, potentially accelerating the search process. Furthermore, LLMs are being explored for automating data science tasks more broadly, including aspects of model selection and pipeline construction (Tiriat et al., 2024). The use of LLMs extends to generating code or configurations for AutoML tools, effectively acting as a meta-learner that understands the relationships between different components. This capability is exemplified by "AutoML-GPT," which directly utilizes an LLM to perform automated machine learning tasks (Tsai et al., 2023). The potential for LLMs to contribute to "green" or more efficient AutoML processes is also being investigated, focusing on reducing computational resources (Tornede et al., 2024).

6.2 State-of-the-art speech-to-text models

Current advancements in Automatic Speech Recognition (ASR) systems are largely driven by sophisticated pre-training techniques that leverage vast amounts of unlabeled data, moving towards more unified and versatile models. Historically, most self-supervised learning approaches for speech focused solely on pre-training an encoder model (Wu et al., 2023). This meant that for generative tasks like ASR, where a sequential decoder is needed, the decoder was often randomly initialized or borrowed from a pre-trained NLP decoder, requiring task-specific supervised data for training (Wu et al., 2023).

A significant leap in this field is Wav2Seq, introduced as the first self-supervised approach to jointly pre-train both the encoder and decoder parts of models for speech data (Wu et al., 2023). Wav2Seq is a self-supervised approach that pre-trains both encoder and decoder parts of speech models, overcoming the limitations of encoder-only pre-training. It creates a "pseudo language" from speech by extracting, pooling and discretizing hidden features into "pseudo characters," which are then deduplicated and tokenized into "pseudo subwords." This "pseudo speech recognition task" significantly improves performance, especially with limited labeled audio data (under 10 hours), closing the gap between encoder-decoder and CTC models in low-resource ASR. As a low-cost second-stage pre-training, Wav2Seq enhances models with

existing pre-trained encoders (e.g., HuBERT). Its versatility allows it to generalize to various architectures, like Seq2Seq and Transducer models, leading to meaningful pattern learning from limited labels and reduced Word Error Rates (WER).

Beyond solely pre-training with speech data, recent advancements integrate text data and multitask learning to create more comprehensive speech-to-text models. Mu2SLAM (Y. Cheng et al., 2023) is a notable example, presenting itself as a multitask, multilingual sequence-to-sequence model pre-trained jointly on unlabeled speech, unlabeled text and supervised data spanning Automatic Speech Recognition (ASR), Automatic Speech Translation (AST) and Machine Translation (MT) in over 100 languages. Mu2SLAM unifies pre-training losses, employing a masked language modeling (MLM) objective on the encoder and a T5-like masked denoising objective on the decoder for unlabeled speech and text. For labeled data, it uses sequence-to-sequence and alignment losses. The model leverages a quantized representation of speech as a target, often utilizing a pre-trained speech tokenizer. A key principle of Mu2SLAM is minimizing modality-specific layers to enforce representation sharing between speech and text, promoting a truly unified model. It also introduces gradual fine-tuning and noisy fine-tuning to bridge the gap between pre-training and fine-tuning, further boosting performance. On ASR tasks, Mu2SLAM has shown competitive results, matching the performance of models fine-tuned with RNN-T decoders despite using a Transformer decoder, which is generally considered weaker for ASR.

Another significant advancement, JOIST (JOint Speech and Text Streaming model) (Sainath et al., 2023), specifically targets streaming ASR models. Unlike traditional approaches that rely on pre-training followed by fine-tuning, JOIST explores joint training with both speech-text paired inputs and text-only unpaired inputs. This direct joint training helps address the challenge of rare-word recognition without the need for large external language models or complex rescoring techniques, which can be computationally expensive or memory-intensive for on-device applications. JOIST injects text by up-sampling text representations (either word-pieces or phonemes) using various duration modeling schemes (fixed repetition, random repetition, sub-word distribution, or align+sub-word distribution). This helps align the text with the temporal nature of speech. The model can also be optimized using the Minimum Word Error Rate (MWER) criterion on unpaired text data, a novel formulation that further improves ASR performance. Crucially, JOIST is designed to maintain streaming capabilities with low latency, measured by metrics like endpointer latency (EP50, EP90) and partial latency (PR50, PR90) and minimize screen flickering as measured by Prefetch Hit Rate (PFHR). It consistently achieves significant WER improvements on rare-word test sets, demonstrating its effectiveness even with large-scale supervised training data.

Further pushing the boundaries of unified models, VoxLM (Maiti et al., 2024) proposes a decoder-only language model that can perform multiple tasks, including speech recognition, speech synthesis, text generation and speech continuation. This approach simplifies multitask integration compared to traditional encoder-decoder architectures that often require task- and modality-specific components. VoxLM integrates discrete speech tokens (derived from self-supervised speech features like HuBERT using k-means clustering, forming "semantic tokens") with the text vocabulary into a shared "VoxLM vocabulary". It uses special tokens (e.g., <start-speech>, <generate-text>) to guide the model for specific tasks like ASR. VoxLM can be initialized with pretrained text LMs (such as OPT) to achieve better performance and faster convergence. The framework demonstrates that ASR can be effectively modeled as a



language modeling task within a joint speech-text framework and it shows improvements in ASR performance, particularly with larger model sizes and increased supervised data.

In conclusion, current advancements in speech-to-text are characterized by a strong move towards self-supervised learning, joint encoder-decoder pre-training and multi-modal integration of speech and text data. Models like Wav2Seq(Wu et al., 2023) demonstrate the power of synthesizing pseudo languages to pre-train all parts of a network, especially under low-resource conditions. Meanwhile, unified models such as Mu2SLAM (Y. Cheng et al., 2023), JOIST (Sainath et al., 2023) and VoxLM (Maiti et al., 2024) are pioneering joint training across multiple modalities and tasks, including streaming ASR, rare-word recognition and even multi-lingual capabilities. These innovations simplify training procedures, improve efficiency, enhance performance on challenging scenarios like limited data or rare words and pave the way for more flexible and general-purpose spoken language models. Future work continues to explore the integration of external language models, improvements in zero-shot capabilities and further consolidation of speech-to-speech and speech-to-text tasks within a single, coherent framework (Y. Cheng et al., 2023; Wu et al., 2023).

6.3 Challenges in speech recognition and NLP systems

The NLP is rapidly transforming how we interact with technology and information. However, this progress brings forth significant challenges that demand careful consideration. These challenges span bias amplification, privacy concerns, hallucination, misinformation generation, accountability issues and the potential for malicious use.

NLP models are trained on vast datasets, often reflecting existing societal biases related to gender, race, socioeconomic status and other sensitive attributes (Garg et al., 2018). Consequently, these biases can be amplified by the model, leading to unfair or discriminatory outcomes (Mehrabi et al., 2021). For example, a résumé-screening tool trained on historically biased hiring data might systematically disadvantage female candidates (Dastin, 2018). Similarly, sentiment analysis models have been shown to exhibit different performance across demographic groups, potentially misinterpreting text written by certain communities (Kiritchenko et al., 2018). Mitigation strategies include careful dataset curation, bias detection techniques during model development and fairness-aware algorithms (J. Chen et al., 2019).

NLP systems often require access to vast amounts of personal data to function effectively. This raises significant privacy concerns, particularly regarding the collection, storage and use of sensitive information. Furthermore, models can inadvertently memorize and reproduce private data from their training sets, a phenomenon known as "memorization attacks" (Carlini et al., 2021). Techniques like differential privacy (Abadi et al., 2016) aim to protect individual privacy while still allowing for useful model training, but often at the cost of reduced accuracy. The increasing deployment of LLMs in cloud environments further exacerbates these concerns, requiring robust data security measures and adherence to regulations like GDPR.

The ability of NLP models to generate human-quality text presents a powerful tool for creating and disseminating misinformation. Hallucinations, where LLMs produce false or misleading information, are a growing concern (Ji et al., 2022). This can be exploited for malicious purposes, such as generating fake news articles (Kozik et al., 2024), impersonating individuals online, or launching targeted disinformation campaigns. Detecting and combating AI-generated misinformation requires developing robust detection techniques and promoting media literacy among the public.

Many state-of-the-art NLP models, particularly LLMs, are black boxes - their internal workings are opaque and difficult to understand. This lack of transparency makes it challenging to identify sources of bias, debug errors, or ensure accountability. Explainable AI (XAI) techniques aim to provide insights into model decision-making processes (Madsen et al., 2022) and remain an active area of research. Without explainability, it is difficult to trust and responsibly deploy these powerful technologies.

The substantial energy consumption and considerable footprint required in the training of large NLP models (Strubell et al., 2019) raise ethical concerns about environmental sustainability and the equitable distribution of resources. Developing more efficient algorithms and hardware, as well as exploring techniques like model compression and knowledge distillation, are crucial steps towards mitigating this impact.

6.4 Identifying ethical concerns and gaps when interacting with a system through natural language and speech

Advancements in NLP are enabling machines to seemingly understand, interpret and even generate human language with continuously enhancing sophistication. However, with these emerging technologies becoming omnipresent and integral part of societal functioning, numerous ethical dilemmas and potential issues emerge. This subsection scrutinizes the multifaceted ethical landscape of NLP, challenges and ethical dilemmas associated with diverse applications, mainly issues concerning algorithmic bias. The next subsection discusses possible mitigation strategies for these concerns.

Ethics in NLP may be understood as an integral field in AI ethics that encompasses the application of ethical principles and values that guide the design, development and deployment of NLP systems in a manner that respects safety, privacy, human autonomy, intellectual property and equity and avoids harmful consequences (Ma, 2023). The scope of AI ethics in NLP is fairly broad as this subfield, inter alia, covers critical areas such as algorithmic bias, fairness in outcomes, accountability for system actions, transparency in decision-making processes, protection of user privacy and the overarching regulatory frameworks governing these technologies (Bender et al., 2021; Ma, 2023). With fast-paced adoption of NLP systems and their integration into daily lives, these ethical principles are becoming a sine qua non for responsible AI models.

Therefore, some of the most pressing ethical issues in NLP may be understood as re-interpretation of a discussion present on a more general level in AI ethics, such as the problem of operationalisation of ethical principles and values in practice, the effectivity of such operationalisation or the question if the AI practitioners are prepared to face moral challenges and dilemmas in their work. This difficulty is well illustrated by the Collingridge dilemma, which highlights a central tension in technology governance: in the early stages of development, it is easier to shape and regulate a technology, but its full consequences are not yet known; later, once the impacts become clearer, the technology is often too entrenched to change easily (Collingridge, 1982). As a result, embedding ethics into NLP systems requires proactive, anticipatory approaches that adapt over time.



Frameworks such as the Ethics Guidelines for Trustworthy AI (European Commission, 2019)³⁰ and Value Sensitive Design (Friedman et al., 2013) offer structured methodologies to address these challenges. They advocate for principles like human agency, fairness, transparency and accountability to be integrated throughout the entire development lifecycle, helping to align NLP systems more closely with societal values.

Another important issue concerns the "dual use" potential of NLP technologies. It refers to the phenomena where systems developed with good intentions can be misused for harmful purposes (Kaffee et al., 2023). Examples include the generation and spread of disinformation, the creation of fake online profiles for malicious campaigns, writing malware, or assisting in fraudulent activities. Even if NLP tools are created with neutral or positive intentions, they can be repurposed for malicious ends. Already marginalized communities are impacted the most by this phenomenon. They often possess less societal power, have fewer resources to protect themselves and are frequently already the targets of discrimination or oppression. Therefore, the negative impact of dual-use NLP is likely to disproportionately affect these communities (Kaffee et al., 2023).

As previously stated, the ethical issues inherent in NLP do not exist in a vacuum. They are not only profoundly rooted in the area of AI ethics, but also intertwined with the evolving landscape of cybersecurity threats. The very properties that make modern NLP models powerful also render them vulnerable to new attack vectors. A primary vector is the manipulation of the model's input-output behaviour through adversarial attacks (Goodfellow et al., 2015). For example, prompt injection attacks aim to bypass safety controls to generate prohibited content, while model evasion techniques manipulate inputs to trigger misclassification, such as fooling content filters (Perez & Ribeiro, 2022). Data poisoning compromises model integrity by inserting malicious examples into training data, creating hidden backdoors or reinforcing specific biases (Carlini et al., 2021). Additionally, models risk data leakage, potentially exposing sensitive personal or proprietary information memorized during training (Carlini et al., 2021). The ability of LLMs to generate persuasive, tailored text also enables large-scale automated social engineering, including phishing and disinformation campaigns that evade detection (Weidinger et al., 2021). These vulnerabilities underscore how insufficient safeguards can turn a model from merely biased to actively unsafe.

Natural language interfaces and speech-based interactions also introduce a new set of vulnerabilities and privacy issues. The simulation of human-like conversation can lower user inhibitions, potentially leading to the inadvertent disclosure of sensitive information (Ali et al., 2025). Furthermore, voice data is not merely a medium for communication. It is a unique biometric identifier that can reveal a speaker's emotional state, potential health conditions and identity. This makes its collection and processing exceptionally sensitive. Moreover, we could observe a risk of voice cloning from insufficiently secured voiceprints, which can lead to sophisticated fraud, impersonation, or even the spread of misinformation. However, ethical concerns in this context are wider. They include issues related to user autonomy, the psychological and social impacts of human-AI relationships, accountability for AI-driven actions and the ownership of co-created content. In addition, data security and privacy are absolutely crucial given the sensitive nature of the information shared with conversational AI tools, voice assistants and other AI interaction tools. For that reason, mitigation of risks requires secure

³⁰<https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1/language-en>



data handling and storage protocols, robust encryption methods and secure data access controls.

NLP systems can have considerable societal influence, affecting individual autonomy, patterns of human interaction and the development of essential skills. They also pose the risk of perpetuating harmful gender stereotypes (Pikuliak et al., 2024) or systematically excluding certain groups from their benefits. The design of user interactions with NLP systems must carefully consider how users perceive and engage with these technologies. Issues such as AI anthropomorphisation, automation bias or AI sycophancy may lead to emotional dependence on AI companions or the risk of users being subtly manipulated by persuasive language technologies. These issues propose growing areas of concern and should be analysed carefully especially considering the vulnerable groups of stakeholders.

The presence of various forms of algorithmic bias in natural language processing is a complex issue that can undermine the benefits of these technologies. It arises when NLP models produce systematically unfair outputs across different groups or reinforce harmful stereotypes (Y. Guo et al., 2024). Bias can originate at various stages—training data, annotation, input representation, model design or broader research decisions (Y. Guo et al., 2024; Hovy & Prabhumoye, 2021).

One of the most widely recognized sources of unfairness is training data bias (Hovy & Prabhumoye, 2021). NLP models learn from large datasets and if these datasets reflect historical prejudices, inequalities, or other systemic imbalances, the models are likely to reproduce them. Several types of data biases can be further discerned.

One form of this is historical bias, where models inherit outdated prejudices, such as racism or rigid gender roles, embedded in past data. Closely related is societal bias, which stems from stereotypes and false assumptions present in the data, causing models to reinforce harmful associations or unfairly limit how certain groups are represented. In addition, selection bias can occur when the data used is not representative of the broader population - models trained primarily on text from specific demographics or cultures may struggle or behave unfairly when applied elsewhere. Human influence also plays a role: annotation bias emerges when annotators' personal beliefs or cultural backgrounds shape how data is labeled, especially in subjective tasks like sentiment analysis. Finally, measurement bias can result from flaws in how data is collected or defined, when tools or criteria used to gather data are not equally suited to all groups, they can systematically disadvantage certain populations.

Bias in natural language processing can originate not just from data, but also from the models and algorithms themselves. Architectural bias arises from design choices in a model's structure—such as layer types or information flow—that can make some architectures more prone to picking up spurious correlations aligned with societal biases (Hovy & Prabhumoye, 2021; Y. Guo et al., 2024). Similarly, loss functions that focus solely on optimizing for accuracy or efficiency, without fairness considerations, can lead to models that perform well on average but consistently disadvantage certain groups (Hovy & Prabhumoye, 2021). Bias can also surface during evaluation. If the metrics or test datasets used to assess models are themselves biased or not representative of diverse populations, evaluation bias can occur. In particular, aggregate metrics like overall accuracy may hide poor performance on minority subgroups, giving a misleading sense of fairness. Finally, interaction bias can develop after deployment. When models adapt to user input, they risk incorporating biased language or behaviour from

those interactions. Over time, this can create feedback loops that reinforce and deepen harmful patterns.

6.5 Mitigation strategies for ethical concerns and future directions

As previously mentioned, NLP systems can have considerable societal influence, affecting individual autonomy, patterns of human interaction and the development of essential skills. They also pose the risk of perpetuating harmful stereotypes or systematically excluding certain groups from their benefits. The design of user interactions with NLP systems must carefully consider how users perceive and engage with these technologies. Issues such as the potential for emotional dependence on AI companions or the risk of users being subtly manipulated by persuasive language technologies are growing areas of concern. The trustworthy NLP systems should be aligned with all three levels of trustworthy AI, legal, security and ethical at the same time (European Commission, 2019).

As for addressing the profound ethical and cybersecurity risks inherent in modern NLP a comprehensive, multi-layered framework is required - a strategy that integrates technical controls, robust governance procedures and user-centric education. Technical strategies may be coded directly into the architecture and operational protocols of NLP systems. They consist of several activities that ought to be taken into consideration and integrated into the technology.

This entails also the rigorous filtering and validation of data, *inter alia*, both before it is processed by a model and before it is presented to a user. On the input side, this is a primary defence against prompt injection attacks, which seek to subvert safety filters. On the output side, it ensures the model does not engage in data leakage or generate overtly malicious content, such as malware code or phishing emails (Perez & Ribeiro, 2022). The training for robustness of AI models also involves exposing a model during its training phase to a curated corpus of adversarial examples designed to simulate real-world attacks. By learning to correctly classify these manipulative inputs, the model develops greater robustness against evasion attacks and targeted misuse (Goodfellow et al., 2014).

As for the mitigation of privacy issues, the introduction of carefully calibrated statistical noise into datasets or model outputs would make it computationally infeasible to determine whether any specific individual's data was included in the training set. This provides a strong, provable guarantee of privacy, directly mitigating the cybersecurity risks of membership inference and model inversion attacks (Dwork & Roth, 2014).

The use of watermarks that embed a computationally verifiable signal into AI-generated content may enhance transparency and human control. These watermarks allow for the subsequent attribution of content to a specific model, serving as an indispensable tool in combating disinformation campaigns and sophisticated fraud leveraging deepfakes, including voice cloning (Kirchenbauer et al., 2024). On the other hand, procedural and governance strategies that refer to policies and processes should maintain human oversight governing NLP development and deployment.

Protecting the entire data supply chain in every single phase of model development and deployment is crucial, too. This requires robust end-to-end encryption, strict access controls and the use of secure hardware, such as trusted execution environments for processing sensitive data. This is a critical defence against data poisoning attacks, wherein an adversary manipulates training data to compromise a model's integrity (Wallace et al., 2021). Red

teaming and adversarial testing involve the proactive and most importantly authorized testing of systems by specialized teams who simulate real-world attackers to discover novel vulnerabilities a priori, before they can be exploited maliciously, or the teams' tests systems against biases (e.g. bias bounty programs). This is a foundation of modern cybersecurity, applied *mutatis mutandis* to the unique attack vectors of NLP models. Rather than releasing powerful models to the public, a staged deployment strategy (e.g., limited-access APIs) allows for controlled testing and learning in a real-world environment. This must be coupled with a formal responsible disclosure policy that provides a secure channel for independent security researchers to report vulnerabilities.

The integration of security and ethical reviews (ethics-based risk and impact assessments) at every stage of the development lifecycle is essential. For NLP, this specifically includes rigorous vetting of the origins and integrity of third-party datasets and pre-trained models, which can themselves be vectors for supply chain attacks. In addition, these assessments evaluate and foresee potential ethical risks that might arise during the development and deployment stage.

However, only an educated and aware society can be resilient against social engineering attacks or misuse of NLP technologies. Public education initiatives and awareness campaigns are critical for teaching citizens to identify AI-driven phishing attempts, recognize deepfakes and apply critical thinking to the information they consume. A literate society will not accept dangerous AI experiments conducted by private organizations or governments on them.

From the developers' and deployers' perspective, it is an ethical imperative that user interfaces clearly and unambiguously disclose when a user is interacting with an AI and when content is AI-generated. This requirement which is also in accordance with the transparency requirements of the AI Act allows users to properly calibrate their trust and vigilance, which in return reduces the efficiency of impersonation and deception campaigns.

The role of legal frameworks, such as the AI Act, proposes a necessary foundation for the effective development and deployment of trustworthy AI systems. They are commencing a fundamental change from voluntary ethical principles to legally binding obligations. By classifying AI systems according to risk and imposing strict requirements for robustness, transparency and bias mitigation, especially on high-risk AI systems, such regulations create powerful incentives for organizations to invest in security and ethical design. The adoption of industry standards, fully harmonized with AI legislation further helps to codify and harmonize best industrial practices.

The trajectory of NLP is one of ever-increasing capability and societal integration. However, this progress is linked to a complex and expanding set of ethical and cybersecurity challenges. The issues of dual use, data privacy and algorithmic bias are not peripheral concerns but are central to the trustworthiness of these technologies. Mitigating these risks is not achievable through any single technical solution. It demands a sophisticated, multi-layered and ongoing commitment. The path forward requires a synthesis of robust technical measures, principled governance, user empowerment through education and clear regulatory oversight. For researchers and practitioners in this field, this endeavour must evolve beyond the mere pursuit of performance metrics to embrace a deeper responsibility for the societal impact of their creations.



7 Exploring bias and ethics in knowledge graph generation

The generation of knowledge graphs involves the systematic construction of structured representations that illustrate entities, their attributes and the relationships between them. These graphs serve as an essential tool for organising large volumes of data (expected from ETD-Hub), making it more accessible and interpretable for both humans and machines. Knowledge graphs have emerged as key components in a variety of fields such as NLP, machine learning and AI. They are employed in applications ranging from search engines and recommendation systems to automated decision-making tools. They are increasingly being leveraged to support AI-driven systems by providing a rich foundation of structured knowledge for tasks that require reasoning and inference.

However, while knowledge graphs present significant opportunities for enhancing AI models, they also introduce a variety of challenges, especially related to bias and ethics. Bias in the generation of knowledge graphs refers to unintentional but systematic distortions in the data, which can arise from skewed or unrepresentative data sources, flawed algorithms, or subjective interpretations during the construction of relationships between entities. These biases can lead to knowledge graphs that reinforce harmful stereotypes, marginalise underrepresented groups, or propagate skewed representations of reality, resulting in models that are less accurate or unfair. As noted by Bolukbasi et al. (2016), even small biases introduced during graph generation can have amplified effects when used to inform machine learning models, affecting the outcomes of AI systems.

Ethics in knowledge graph generation, on the other hand, concerns the responsibility of creating these graphs in ways that prioritise fairness, transparency and accountability. The ethical challenges are particularly pertinent when knowledge graphs influence systems that guide decision-making especially in high-stakes environments. In the context of the ALFIE project, the knowledge graph will be integral in ensuring that AI models created by and for users align with ethical guidelines and relevant laws. Furthermore, the ethical implications of knowledge graphs extend to data security, particularly when the graph integrates sensitive data, such as demographic data, which must be carefully managed to prevent misuse or breaches of confidentiality (Binns, 2017).

Failure to adequately address bias and ethical considerations during knowledge graph creation can lead to systems that perpetuate existing societal inequalities. For instance, biased AI models in software development can result in tools that reinforce stereotypes or overlook underrepresented groups. One example is biased recommendation systems, where algorithms trained on non-representative data can perpetuate homogenous content or products, thereby limiting diversity in user recommendations and creating echo chambers. Additionally, AI models used in bug detection or code review systems might be less effective at identifying issues in code written by developers from diverse backgrounds if the training data predominantly reflects code from a narrow demographic or specific programming languages. This can lead to skewed or incomplete insights, potentially disadvantaging certain developers or communities. Moreover, if the datasets used to train these systems are not diverse or inclusive, AI models could reinforce existing biases in programming practices or algorithmic decision-making, thereby perpetuating inequalities in the software engineering field (Memarian and Doleck, 2023). These examples highlight the importance of ensuring fairness,

transparency and inclusivity in AI model development, especially in the context of software engineering, where the consequences of biased systems can affect both the industry and society at large.

As knowledge graphs continue to inform key decisions in AI-powered systems, it becomes increasingly vital to ensure that the creation of these graphs adheres to ethical principles. Early attention to the mitigation of bias and adherence to ethical standards during the development of knowledge graphs can help establish more transparent, accountable and reliable AI systems. Furthermore, as Binns (2017) emphasises, an ethical approach to knowledge graph creation can help align AI systems with societal values, fostering trust and improving the social acceptability of these technologies. Thus, the exploration of bias and ethics in knowledge graph generation is not merely an academic exercise; it is foundational to the development of AI systems that are both technically effective and ethically responsible.

In the following subsections, we will explore the state-of-the-art tools used in knowledge graph generation (Section 7.1), examine the specific challenges related to bias and ethical considerations (Section 7.2) and discuss the strategies that can be employed to mitigate bias during the construction of knowledge graphs (Section 7.3). These discussions are crucial to understanding how to create knowledge graphs that serve as ethical and reliable resources for AI model development which is the core of the ALFIE project.

7.1 State-of-the-art tool for creating knowledge graphs

The development of knowledge graphs has been significantly impacted by the rise of advanced computational tools and technologies that automate, enhance and optimize their creation. These tools leverage various approaches, from manual data curation and database management systems to more sophisticated AI-based solutions. As knowledge graphs are being adopted across different industries, such as healthcare, finance and social media, a growing array of tools have emerged which support both the creation and the efficient management of these complex structures.

7.1.1 Traditional database and graph management systems

Historically, knowledge graphs were created using relational databases and basic graph management systems (Perez et al., 2018). These early systems were constrained by their inability to efficiently process unstructured data, which made it difficult to scale up knowledge graphs to the size needed for modern applications. However, they laid the groundwork for more advanced systems. Resource Description Framework (RDF) (Brickley and Guha, 2014) and SPARQL (Garlik, 2013) were among the earliest technologies used to represent and query knowledge graphs, forming the backbone of early semantic web tools. These systems were highly structured, requiring specific knowledge about data schemas to work effectively, which made them limited in flexibility. Despite these limitations, tools like Apache Jena³¹ and Virtuoso³² laid the foundations for what we know today as graph databases, making it easier to query and navigate relationships within structured datasets.

³¹ <https://jena.apache.org/>

³² <https://virtuoso.openlinksw.com/>

7.1.2 The emergence of graph databases

A more robust and flexible solution emerged with the development of graph databases. These systems, such as Neo4j³³, ArangoDB³⁴ and JanusGraph³⁵, represent data in terms of nodes (entities) and edges (relationships), making it easier to capture the inherent interconnectedness of data. Neo4j remains one of the most popular graph databases, offering both a powerful query language, Cypher and a rich set of features for creating and managing complex relationships (Robinson et al., 2015). Neo4j has found widespread use in various industries, including social network analysis, fraud detection and recommendation engines, by offering high-performance graph traversal and visualisation (Hogan et al., 2020). However, despite its capabilities, Neo4j and other graph databases face significant limitations in scaling across vast datasets, especially those with more than a few billion nodes or edges. As these databases rely on a single-node architecture for processing, performance can degrade significantly when handling large-scale graph data, resulting in slower query responses and increased latency. To manage very large graphs, it is often necessary to distribute the graph across multiple machines, but doing so introduces complexities in data consistency and query processing, limiting the scalability of these systems for massive datasets or real-time applications (Robinson et al., 2015). ArangoDB and JanusGraph contribute to the field by allowing distributed graph databases, which can store vast amounts of data across multiple servers, ensuring horizontal scalability and robustness for big data applications. These databases, however, introduce the challenge of maintaining consistency in distributed systems, which can affect the reliability of results (Fellows et al., 2021). The emergence of distributed graph computing frameworks such as Apache TinkerPop³⁶ has been a critical step forward, enabling users to run graph algorithms at scale across cloud-based environments. TinkerPop is an open-source graph computing framework that allows developers to apply graph traversal and analysis techniques across various graph databases, which is crucial when dealing with large datasets across heterogeneous systems (J. Cheng et al., 2020). TinkerPop's ability to integrate with different graph databases (like Cassandra, HBase and Neo4j) enables organisations to move towards a more flexible, scalable and integrated approach for managing knowledge graphs at scale.

7.1.3 AI-driven tools and NLP integration

Alongside the growth of graph databases, there has been a significant leap forward in AI-driven knowledge graph generation. Tools now incorporate sophisticated ML algorithms and NLP models to automate the extraction of knowledge from unstructured sources, such as text documents, web pages and scientific literature. SpaCy (Honnibal et al., 2020) and Stanford NLP (Manning et al., 2014) are two highly recognised NLP libraries for processing textual data to identify named entities, relationships and key concepts, forming the basis for knowledge graph construction. For instance, SpaCy has recently incorporated deep learning models that can extract more nuanced entity relationships from text, which is critical for generating large, complex knowledge graphs from raw documents. In addition, SpaCy offers various pre-trained models for tasks like dependency parsing and text classification, which are particularly useful in identifying connections and organising entities into structured knowledge. On the other hand,

³³ <https://neo4j.com/>

³⁴ <https://arangodb.com/>

³⁵ <https://janusgraph.org/>

³⁶ <https://tinkerpop.apache.org/>

Stanford NLP provides a comprehensive suite of tools for natural language processing, including state-of-the-art named entity recognition (NER), part-of-speech tagging and coreference resolution, which are instrumental in building relationships between entities in knowledge graphs. These tools allow for the extraction of rich, semantic data from large text corpora, which can then be transformed into knowledge graphs that are both accurate and semantically meaningful.

Likewise, tools like Open Information Extraction (OpenIE), developed by the University of Washington (Etzioni et al., 2008) use ML techniques to extract triples (subject-predicate-object) from unstructured text, thus facilitating the creation of knowledge graphs directly from text. OpenIE's advantage lies in its ability to automate knowledge graph generation by processing natural language text without requiring a predefined ontology, making it more flexible for a variety of domains. However, OpenIE faces challenges in ensuring the extraction of high-quality triples in complex texts, where context and ambiguity play a significant role. More recent approaches, such as BERT-based OpenIE (using the transformer model BERT for deeper contextual analysis), have been developed to address these concerns by leveraging contextual embeddings to extract more accurate relationships and handle polysemy (Tang et al., 2021).

Another advanced tool, Relational Graph Convolutional Networks (R-GCNs), integrates graph neural networks with knowledge graph creation. These systems use graph-based learning techniques to automatically identify patterns in the data, iteratively improving the quality of the generated graphs. They are particularly useful in applications requiring the integration of multiple data sources, such as in multimodal applications that combine textual, visual and structured data sources for knowledge graph creation (Schlichtkrull et al., 2018). These tools are being used in fields like recommendation systems and knowledge discovery, where graph-based representations can reveal hidden relationships within large and diverse datasets.

7.1.4 Future directions and challenges

While the current state-of-the-art tools for creating knowledge graphs have been transformative, the field continues to evolve. The integration of semantic reasoning with knowledge graphs is an area of active research. For example, Web Ontology Language (OWL) and RDFS allow developers to create ontologies that add semantic layers to graphs, enabling machines to reason about the data. This opens the possibility for more sophisticated AI systems that can infer new knowledge from existing relationships in the graph. However, the complexity of integrating semantics at scale remains a challenge.

Additionally, scalability continues to be a critical concern. While graph databases and distributed computing frameworks have made strides, handling graph data at the scale of the entire web or across billions of interconnected entities requires improvements in processing power, distributed storage systems and algorithm optimization. Further work is needed to enable seamless integration of knowledge graphs with real-time data streams, which will be essential for future applications like autonomous systems or real-time decision-making tools.

7.2 Challenges in bias and ethical consideration in knowledge graphs

As knowledge graph technologies continue to mature, several critical challenges regarding bias and ethics must be addressed. These challenges are compounded by the increasing use of knowledge graphs in sensitive and high-stakes applications, such as law enforcement,



healthcare and social media moderation. The potential for harmful consequences due to bias in knowledge graphs has prompted growing concern over how to ensure that these tools are both effective and fair.

7.2.1 Data bias and representation bias

One of the most significant sources of bias in knowledge graph generation stems from the data used to build them. Knowledge graphs are often constructed from large-scale datasets scraped from the web, social media platforms, discussion forums or corporate databases. These sources, while abundant, are not always representative of the full diversity of human experience. For example, datasets scraped from English-language websites often underrepresent non English-speaking communities, creating knowledge graphs that skew towards Western perspectives and ignore or marginalise other cultures (Binns, 2017). The data quality issue is particularly pronounced when the data used to generate knowledge graphs is uncured or sourced from biased platforms. For instance, social media platforms like Twitter or Facebook often reflect the biases inherent in their user bases, leading to the amplification of certain viewpoints while marginalising others (Bolukbasi et al., 2016; Voit and Paulheim, 2021). Similarly, news outlets may perpetuate bias based on their editorial stance, resulting in knowledge graphs that amplify certain narratives while neglecting others (Kraft and Usbeck, 2022). This is particularly relevant to the ALFIE project which intends to develop knowledge graph(s) based on information obtained from the ETD-Hub, a forum-like platform for engaging experts on AI ethics related subjects. Representation bias is another critical concern. Knowledge graphs are constructed from both structured and unstructured data, but the relationships identified in the graph often rely on the availability and quality of the underlying data. In domains where the data is sparse, such as in rare diseases or underrepresented demographics, knowledge graphs may fail to capture critical information or may produce imbalanced relationships, thereby reinforcing historical disparities. In fact, this may be the case for ETD-Hub where data is likely to be sparse with multiple threads discussing different but related topics, thus requiring further processing and alignment to map entities to multiple knowledge graphs (Zhu et al., 2024).

7.2.2 Ethical and privacy concerns

The ethical challenges surrounding knowledge graphs also involve privacy concerns. Many knowledge graphs are created by aggregating vast amounts of personal data from sources such as social media, online reviews, or customer transactions. The aggregation of this data can expose individuals to privacy violations if proper safeguards are not in place. For instance, knowledge graphs that map social connections or personal behaviours may inadvertently disclose sensitive information about individuals, even if the data was originally collected in a seemingly anonymous form (Dastin, 2018).

Moreover, as knowledge graphs are increasingly used in decision-making tools, such as automated hiring systems, loan approval processes, criminal risk assessments or in ALFIE's case, software development, the ethical consequences of biased graphs become even more significant. These systems, if based on biased knowledge graphs, could perpetuate systemic discrimination, such as racial or gender biases, leading to unfair or discriminatory outcomes. For example, AI tools used in the criminal justice system, such as risk assessment algorithms, have been shown to be biased against minority populations, often due to biased training data

(O'Neil, 2016). Knowledge graphs used in such systems must be scrutinised to ensure they do not perpetuate or exacerbate these issues.

7.2.3 Lack of transparency and accountability

Another pressing ethical issue in knowledge graph generation is the lack of transparency in how knowledge graphs are built and the algorithms that underpin them. In many instances, the processes that extract relationships and entities from data are opaque, making it difficult to understand how biases are introduced. This lack of transparency can lead to issues of accountability, particularly in high-stakes applications where decisions based on knowledge graphs have significant societal impacts. For example, automated hiring systems that rely on knowledge graphs to match candidates to job openings can perpetuate biases based on factors such as gender or race. If the knowledge graph is built using biased data sources or flawed algorithms, the resulting system may inadvertently disadvantage certain demographic groups. Without transparency in the creation of these knowledge graphs, it is challenging to identify and correct such biases.

7.2.4 Directions for addressing bias and ethics

The development of XAI is one potential direction for improving transparency and addressing ethical concerns in knowledge graph generation (Tiddi et al., 2020). XAI aims to make AI systems more interpretable by providing clear explanations of how decisions are made. By applying XAI principles to knowledge graph generation, developers can create systems that allow users to understand how knowledge is represented and how relationships between entities are drawn, improving accountability and trust.

Moreover, ethical frameworks such as Fairness, Accountability and Transparency (FAT)³⁷ are gaining traction in AI research as tools for addressing the societal impacts of AI systems, including knowledge graphs. These frameworks advocate for systematic approaches to ensuring that AI technologies are fair and equitable and do not inadvertently reinforce existing inequalities (Memarian & Doleck, 2023). By embedding these principles into knowledge graph generation, developers can mitigate bias and foster more ethical practices in the field.

7.3 Mitigation strategies for limiting bias in knowledge graphs

Addressing bias in knowledge graphs requires proactive strategies to detect, correct and prevent the introduction of biases in the data, algorithms and outcomes. Several methods have been proposed and are currently being explored to mitigate bias and improve the fairness and reliability of knowledge graphs.

7.3.1 Bias Detection and Correction

One of the most direct approaches to mitigating bias in knowledge graphs is the application of bias detection and correction techniques. These techniques involve systematically identifying and addressing biases that may exist in the underlying data sources. Bias detection can be performed through statistical analysis, which involves evaluating the representation of different groups or categories within the graph. For example, researchers can measure how frequently certain entities (e.g., women, ethnic minorities) appear in the graph and whether certain relationships are overrepresented or underrepresented based on these entities (Bolukbasi et

³⁷ <https://www.fatml.org/>

al., 2016). Once detected, corrective actions can be taken to balance the graph, such as introducing underrepresented data or adjusting weights to reflect a more equitable representation.

Another approach is algorithmic fairness methods, which apply fairness constraints to the algorithms that generate knowledge graphs. These constraints are designed to ensure that the graph does not disproportionately favour certain groups or individuals over others. For instance, adversarial debiasing methods can be used to train ML models to minimise biases in the generated knowledge graph while still maintaining high levels of predictive performance (Grari et al., 2023; Cai and Wang, 2018; Arduini et al., 2020; Chuang et al., 2025; Luo et al., 2025).

7.3.2 Data diversification and augmentation

Another essential strategy for mitigating bias is the use of data diversification. By sourcing data from a wide range of domains, cultures and perspectives, developers can ensure that the knowledge graph is more representative of the diversity of human experience. Data augmentation techniques, such as generating synthetic data, can also be employed to fill in gaps where certain perspectives are underrepresented or omitted from the data. These strategies can help create more balanced and inclusive knowledge graphs, reducing the risk of exclusion or bias against marginalised groups (Binns, 2017).

7.3.3 Incorporating ethical frameworks

Ethical frameworks provide a structured approach for ensuring that knowledge graphs are built responsibly and fairly. These frameworks typically include principles such as transparency, accountability and inclusivity, which guide developers in making ethical decisions throughout the graph creation process. Frameworks like FAT (i.e., Fairness, Accountability and Transparency) advocate for auditing and continuous evaluation of AI systems to ensure they adhere to ethical standards (Memarian & Doleck, 2023). Incorporating such frameworks during the design and evaluation of knowledge graphs helps identify potential ethical risks early in the process.

7.3.4 Human-In-The-Loop (HITL) approaches

Incorporating a human-in-the-loop (HITL) approach can also be an effective way to mitigate bias in knowledge graphs. HITL strategies involve human experts reviewing and validating the outputs of automated knowledge graph generation systems to ensure that the resulting graphs are fair and accurate. This is particularly useful when the data or relationships in question are nuanced and require domain expertise to interpret effectively. By combining machine-generated insights with human judgment, HITL approaches can improve the quality and fairness of knowledge graphs, particularly in high-stakes domains (Schröder et al., 2022; Tsaneva et al., 2025).

7.3.5 Privacy-preserving techniques

Given the sensitive nature of much of the data used to generate knowledge graphs, particularly in areas like healthcare and finance, privacy-preserving techniques are essential for ensuring the ethical use of these systems. Differential privacy (Dwork, 2006) and/or federated learning (Peng et al., 2021) are two such techniques that allow knowledge graph generation to occur without compromising the privacy of individuals. Differential privacy adds noise to the data in

a way that prevents the identification of individual users, while federated learning allows models to be trained across decentralised datasets without centralising the data. Both approaches ensure that knowledge graphs can be built in a way that respects users' privacy rights while still benefiting from large, diverse datasets.

7.3.6 Future directions

The future of mitigating bias in knowledge graphs lies in the continued development of more sophisticated bias detection tools, automated fairness techniques and explainable solutions that improve transparency and trustworthiness. Furthermore, the ethical implications of knowledge graphs will continue to be a critical area of research, particularly as these graphs are used in increasingly impactful applications such as autonomous systems, legal decision-making and health diagnostics. The ongoing integration of interdisciplinary insights from ethics, law and social sciences will likely guide the development of more equitable and fair knowledge graphs in the future.



8 Leveraging code generation to generate ethical and unbiased codebases

8.1 Overview of code generation technologies

The underlying technology behind code generation has been around for quite some time. The main limitation was that it was never adequate for generating large blocks of code; however, it was quite effective for tasks such as correcting syntax, suggesting documentation, or proposing small code snippets based on unfamiliar libraries. With recent advances in LLMs, these systems have shifted from rule- or template-based methods to exclusively using transformer architectures. This shift has enabled several innovations, including much larger context windows, improved memory, the ability to explain reasoning to some extent and the capability to generate large sections of code in a single attempt. While there are still many issues with this approach, it has undeniably improved developer productivity over the past few years. Tasks such as combing through documentation have, to a large extent, been replaced by these systems integrated into code editors (Torka & Albayrak, 2024). This is not to say that documentation is unhelpful, rather, these systems enable users to sift through large amounts of documentation and pinpoint the exact information they need far more quickly than before.

The main applications of code generation technologies are in domains such as automated code refactoring, rapid prototyping and, in some cases, complete automation (Jin et al., 2025). While the latter is not always used or preferred, it is possible to an extent. Various model optimization techniques have emerged to distil large models so they can run on devices without specialized hardware. Since these LLMs are usually quite large, running an extremely accurate model locally remains a significant challenge. Nevertheless, given recent advancements, many editors and CI/CD environments now integrate such automation.

A more recent trend, as of the time of writing, is the rise of agentic systems, systems capable of autonomously performing tasks end to end. While much of the discourse around them is driven by hype, it is interesting to consider the possibility of such systems performing an unlimited number of tasks without manual input. That said, current systems do not truly operate this way; many tasks performed by so-called agentic systems have been manually programmed, with the system simply stringing together pre-existing modules in an intelligent manner.

In the context of AutoML, it is indeed possible to automatically generate entire systems and train new models using such agents. However, this is not always the most optimal approach. These systems can generate hyperparameter tuning scripts, data processing steps and even complete machine learning pipelines. While they are not perfect and have notable limitations, they can sometimes produce quite reasonable results. Perhaps with further research, it will one day be possible to fully automate this process without raising legal or ethical concerns, but at present, that is not yet the case.

8.2 Ethical frameworks and challenges in code generation - Mitigation strategies

Code generation systems face several significant challenges, most of which can be traced back to the quality and nature of their training data. In the past, it was not feasible for models



to process such vast quantities of data; now that it is possible, the challenge has shifted to curating the data these models are trained on (Improta et al., 2025). In code generation specifically, there is a notable imbalance in the availability of high-quality code for niche or emerging languages and problem domains. As a result, the code generated by these systems might not be optimal, even if it appears correct at first glance. Another major difficulty lies in identifying copyrighted or biased data without some level of human oversight. While techniques such as normalization, duplicate removal, stripping comments and removing personal information can help ensure cleaner datasets, these methods are far from perfect. For example, there have been numerous cases of leaked API keys caused by insufficient data sanitization (Chen & Jiang, 2024).

The most reliable way to prevent such issues is to curate datasets manually or semi-automatically, ensuring they are free from private content, duplicates and known vulnerabilities (Xu et al., 2025). However, doing so greatly reduces the volume of data available for training, which is why this approach is rarely applied in practice. Another challenge lies in handling edge cases for code generation tasks. While an LLM can often produce functional code, its lack of full context may prevent it from generating robust, production-ready systems in a zero-shot setting. Many edge cases are therefore missed and must be addressed through human intervention. One way to train models to handle these situations is by providing manually annotated datasets containing examples of such edge cases. However, this process is costly and labor-intensive.

Given the risks related to unclear code, licensing issues, copyright violations and attribution concerns, practical policies and updated legal guidelines are essential. Static analysis tools could be used to verify whether generated code follows required protocols, or to detect vulnerabilities, security flaws, or potential legal risks. If any such issues are found, they should be flagged for human review before any further action is taken (Z. Wang, 2025).

AutoML systems in particular amplify the importance of clean, diverse training data, as they often draw from a wide range of sources when generating code. This breadth also increases the risk of using biased models or unsafe preprocessing steps, making it difficult to trust them for fully automated, end-to-end code generation (Gao et al., 2024).

For these reasons, safeguards must be implemented. Any task with the potential to cause harm if executed incorrectly should always have a built-in mechanism for human review before being carried out.

8.3 Privacy concerns in code generation

LLMs are highly proficient at generating content - after all, they were designed to predict the next word or phrase given a context. In practice, they are trained on massive amounts of internet data, enabling them to produce coherent and contextually relevant outputs. However, this also means they can fall into common pitfalls, one of the most serious being the memorization of proprietary or sensitive data without safeguards. To an LLM, all data is essentially the same; it cannot inherently distinguish between public and proprietary content. For a user generating code, this can result in stolen or copyrighted material being unknowingly integrated into their codebase, leading to potential legal consequences.

This leakage of private or sensitive information could be avoided if training datasets excluded such content (Niu et al., 2023). Unfortunately, most large-scale LLMs are trained on datasets



that include it, making prevention challenging. In some cases, sensitive details, such as API keys, may leak, leading to costly breaches. More advanced risks include blind membership inference attacks, where it becomes possible to infer whether specific data was part of the training set. While these attacks are not trivial, since LLMs do not maintain explicit indexes of their data sources, they are not impossible.

To mitigate such risks, GDPR-like compliance frameworks for LLMs would be highly valuable, if not essential (Geng et al., 2025). While such measures are still in development, the responsibility currently falls on the end user to clearly label any code obtained from an LLM. In the case of AutoML, automated feature selection can also inadvertently reveal private information if safeguards are absent. Fully automated processes further complicate the task of identifying and auditing such issues. Therefore, it is critical to incorporate privacy checkpoints, either through human oversight or through rule-based systems that operate independently of generative models.

Research into privacy-preserving learning techniques, such as federated learning and incremental learning, is ongoing. These methods aim to update models with new data while obfuscating or removing identifying information, ensuring that private data cannot be traced back to individuals.

8.4 Future directions and recommendations

While there is a lot of promising research in these domains, there is still a long way to go before we reach optimal performance in code generation. In the case of ALFIE, the modular nature of the system allows the addition of new functionalities as and when state-of-the-art research progresses. One of the major areas currently being explored is improving the robustness and reliability of solutions generated by LLMs. At present, these solutions often work but are not necessarily the most optimal or secure way of addressing certain problems (Chen & Jiang, 2024). As a result, developers may need to invest more effort in making an LLM-generated solution work than they would if they had written it themselves.

A key limitation is that LLMs cannot process context sizes as large as an entire codebase. Even techniques like sliding-window prompts, which attempt to overcome this, often fail to incorporate the full context, leading to omissions in important details. Beyond improving reliability, another area of future work is the training process itself. Currently, most LLMs are trained on data scraped from the internet - data that often ignores ethical guidelines and privacy requirements. While this practice is not entirely legal, it is frequently overlooked due to the usefulness of these systems. A significant direction for future research is therefore the development of LLMs trained entirely on legally sourced and ethically collected datasets.

Another challenge is the risk that LLM-generated code may contain copyrighted content. If used unknowingly, this can lead to legal consequences for the end user. Several mitigation methods exist, but none are foolproof (Jin et al., 2025). One approach uses a separate LLM, trained specifically to detect legal risks, as a post-processing step. While promising, this method is still imperfect. Addressing this issue will require the creation of new ethical and legal guidelines tailored specifically to code generation.

In addition, results could be improved by incorporating community-driven feedback. Currently, most code-generation systems are trained once and updated periodically. This process is memory-intensive, costly and therefore infrequent, meaning systems are often outdated and



fail to incorporate the latest research. Retrieval-Augmented Generation (RAG) is one approach to mitigating this, as it uses the model's existing knowledge to retrieve relevant information from a database of up-to-date documents. While effective in some cases, it is not a universal solution and remains an active research area (Zeng et al., 2024).

For AutoML systems, there is also a need for transparent logs generated during pipeline creation. Such logs would make it possible to audit a system's decisions and identify flaws. Unfortunately, with current technology, this is difficult, as large models often operate as black boxes.

9 ALFIE's Pilot Use cases

9.1 Ethical AI tools for Connected Automated Vehicles (CAVs)

The advent of Connected Automated Vehicles (CAVs) marks a paradigm shift in transportation, integrating advanced automation with vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication to enhance safety, efficiency and mobility. Defined across a spectrum of automation levels by the Society of Automotive Engineers (SAE), CAVs range from partial automation (Level 2) requiring human oversight to full autonomy (Level 5) where human intervention is obsolete³⁸. A cornerstone of their safety promise lies in mitigating human error, which accounts for approximately 90% of road accidents, with driver drowsiness implicated in 10-20% of these³⁹. Drowsiness detection systems employing physiological, behavioural and vehicle-based metrics aim to monitor driver alertness and trigger interventions, particularly in semi-automated CAVs where human-automation collaboration is critical (Sahayadhas et al., 2012).

Machine learning applications in CAVs are extensive, ranging from object detection using Convolutional Neural Networks (CNNs) to analyse visual data, to driver's state monitoring in partially automated systems. Connectivity enhances these capabilities by enabling real-time data sharing, which can improve system performance beyond traditional in-vehicle methods (Díaz-Santos et al., 2024). A key area of interest is safety in partially automated vehicles, where drivers must remain vigilant to intervene when necessary. Here, drowsiness detection exemplifies a critical application, with research leveraging CNNs for facial feature analysis and deep neural networks for physiological signals to achieve high detection accuracy (Ebrahimian et al., 2022; Yaman et al., 2023). These efforts underscore machine learning's role in addressing human factors alongside technical challenges.

Despite its promise, machine learning in CAVs faces hurdles, including model robustness across varied conditions, data privacy concerns and the effective integration of connectivity. Studies highlight the need for tailored approaches in partially automated contexts, such as combining postural and physiological indicators for drowsiness monitoring (Perrotte et al., 2024). This introduction lays the groundwork for a comprehensive literature review on machine learning in CAVs, with a subsequent focus on drowsiness detection as a vital safety mechanism in partially automated systems.

9.1.1 State-of-the-art algorithms for CAVs

Connected Automated Vehicles (CAVs) integrate autonomous driving with vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) connectivity, aiming to enhance safety, efficiency and user experience across various automation levels (SAE 1–5). Machine learning algorithms are pivotal in this domain, enabling CAVs to process complex data for perception, decision-making and driver monitoring. In partially automated systems (SAE levels 2 and 3), where human intervention remains necessary, drowsiness detection emerges as a critical safety application. This section explores the state-of-the-art ML algorithms employed in CAVs, with a particular

³⁸ SAE (2018), Sae international releases updated visual chart for its “levels of driving automation” standard for self-driving vehicles.

³⁹ <https://www.nhtsa.gov/risky-driving/drowsy-driving>

focus on drowsiness detection, synthesizing recent advancements, methodologies and their implications.

Machine learning algorithms underpin numerous CAV functionalities. Convolutional Neural Networks (CNNs) excel in perception tasks, such as object detection and lane recognition, processing visual data from cameras with high accuracy (Yaman et al., 2023). Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are used for temporal data analysis, aiding in trajectory prediction and traffic flow modelling. Reinforcement learning supports autonomous decision making, optimizing navigation in dynamic environments. In fully autonomous CAVs (SAE levels 4–5), these algorithms integrate with connectivity features, leveraging 5G/6G networks for real-time data exchange to enhance situational awareness (Díaz-Santos et al., 2024). However, in partially automated CAVs, machine learning extends beyond vehicle control to driver monitoring, where drowsiness detection is paramount to ensure readiness for manual takeover.

9.1.2 Drowsiness Detection Techniques

Drowsiness detection in CAVs employs a range of ML approaches, tailored to the unique challenge of reduced driver engagement in partially automated modes. These techniques range from image-based analysis, to biological signal analysis and vehicle movement analysis.

To perform an image-based drowsiness detection analysis the first method uses in-vehicle cameras to monitor facial features - eye closure, yawning and head pose. A study by Yaman et al. (2023) demonstrates a real-time system achieving 99% accuracy using CNNs to classify eye states and facial expressions under varied lighting conditions. A further review (Samy Abd et al., 2024) used CNNs alongside Support Vector Machines (SVM) and Random Forests, noting their adaptability to driver monitoring.

Moreover, as mentioned above the second method for drowsiness detection is to utilise biological signals. Biological signals data, such as electroencephalogram (EEG) and electrocardiogram (ECG), provide direct indicators of drowsiness. Several deep neural networks deployed to analyse ECG and respiration signals simultaneously, achieving 77–97% accuracy in classifying drowsiness levels, highlighting the method's precision for early detection (Ebrahimian et al., 2022). Such approaches are less common in CAVs due to sensor invasiveness but remain highly effective.

Finally, Vehicle Movement Analysis analyses behavioural cues to determine drowsiness detection. These cues can range from steering wheel angles and lane-keeping patterns and are analysed using ML models like decision trees or SVMs. A recent research (Perrotte et al., 2024) explored this in a simulator study with 22 participants under SAE level-2 automation, combining postural and physiological indicators to infer drowsiness, though ML was not explicitly applied. This method is less sensitive to early drowsiness but compliments other techniques.

However, these three techniques are not used in isolation of one another. Combining multiple data sources could enhance accuracy. Martin Gjoreski et al. (2020) investigated physiological and visual signals using end-to-end deep learning, identifying emotional activation and facial action units as robust drowsiness predictors. In CAVs, hybrid systems could leverage connectivity to integrate external data, a potential yet underexplored advantage.

9.1.3 Challenges and fairness in CAV systems

9.1.3.1 Technical challenges

A primary technical challenge is the collection of high-quality, diverse data for training machine learning models. Ethical considerations limit the ability to collect data from drivers in states of severe drowsiness while driving, often confining studies to simulated or controlled environments (Yaman et al., 2023).

Moreover, the need for data encompassing varied driving conditions, lighting scenarios and driver demographics to ensure model generalization (Samy Abd et al., 2024) is another technical challenge.

Furthermore, the requirement for real-time detection imposes computational constraints, as models must process data quickly on in-vehicle hardware. Proposing a low-cost driver monitoring system using deep learning emphasizes the need for efficient algorithms to balance accuracy and computational load (Khalil et al., 2025). This challenge is particularly acute in CAVs, where resources are shared across multiple systems.

Another technical factor that CAVs have to settle in is that the driver behaviour and physiological responses to drowsiness vary widely, complicating model development. Personalized models or meta-learning approaches may be necessary to handle individual differences, such as varying signs of fatigue, like eye rubbing or head nodding (Samy Abd et al., 2024). Furthermore, physiological signals for drowsiness detection can differ between individuals, suggesting the need for adaptive algorithms (Atwya & Panoutsos, 2020).

Last but not least the integration of such algorithms with the current CAVs systems is another challenge. Integrating driver monitoring systems with CAV architectures requires ensuring compatibility with existing vehicle functions and effective data sharing. Using 5G/6G connectivity for real-time data transmission, it highlights the need for seamless integration with CAV communication protocols (Díaz-Santos et al., 2024). In addition, leveraging connectivity in CAVs for enhanced monitoring introduces issues like data transmission latency, network reliability and security. It is important to manage these factors to maintain real-time capabilities and trustworthiness, especially when sharing driver state data with other vehicles or infrastructure (Díaz-Santos et al., 2024).

9.1.3.2 Fairness issues

On top of the technical challenges faced by the integration of machine learning algorithms in CAVs there are some challenges concerning the fairness and unbiased nature of these algorithms. Recent research has highlighted the presence of demographic biases in models designed for driver drowsiness detection, a critical function for CAVs. One study examining individual differences in drowsiness detection found that models often perform inconsistently across diverse groups, indicating that a “one-size-fits-all” approach may overlook important variations in physiological and behavioural responses among different demographic segments (X. Wang & Xu, 2016). In a related study on autonomous driving systems, researchers uncovered fairness issues in detection tasks, such as a higher proportion of undetected instances among certain age groups, which, although focused on pedestrian detection, underscores the broader risks of deploying biased models in safety-critical contexts (Xinyue et al., 2025).

Complementing these findings, a survey on dataset biases in drowsiness detection further demonstrated that training data often lack the demographic diversity necessary for robust model performance across all populations. This work argues that models built on non-representative data can inadvertently favour certain groups over others, leading to unequal safety outcomes in real-world scenarios (Fu et al., 2024).

Moreover, recent studies have underscored that the substantial physiological differences between individuals, such as variations in baseline neural activity and the expression of specific EEG features, pose a critical challenge for integrating machine learning models in connected autonomous vehicles for driver drowsiness detection. Recent advancements in drowsiness detection employing EEG-based data demonstrated that cross-subject drowsiness recognition which applies inter-subject variability (e.g., differences in the presence and distribution of Alpha spindles and Theta bursts) significantly degrades model performance if not properly addressed, thereby complicating the creation of calibration-free systems that work reliably across diverse drivers (Cui et al., 2023). Similarly, single-channel EEG data, while promising for non-intrusive monitoring, still suffers from individual specific signal variations that must be managed through model designs that are both compact and interpretable (Cui et al., 2022).

In addition, comprehensive reviews of physiological signals for drowsiness detection reveal that individual variability, stemming from factors such as age, gender and inherent physiological differences, remains a persistent challenge in ensuring the robustness and fairness of these systems. A systematic review of physiological signal-based drowsiness detection methods discusses how these variations can lead to inconsistent detection performance, underscoring the need for diverse, representative datasets and advanced calibration techniques to improve generalization across populations (Saleem et al., 2023).

Last but not least, research indicates that cultural and behavioural differences among drivers present significant challenges when developing machine learning models for drowsiness detection in connected autonomous vehicles. Behavioural cues such as eye blink frequency, facial expressions and head movements can vary markedly across different cultural contexts due to distinct social norms and attitudes toward fatigue. For example, drivers from certain regions might display more subtle signs of drowsiness compared to those from other cultures where overt fatigue behaviours are more common. Such variability means that models trained on data from one cultural group may not generalize well to others, making it essential to incorporate diverse, culturally representative datasets during development (Gwak et al., 2018).

Moreover, behavioural differences related to risk perception, driving style and even the willingness to report fatigue further complicate the detection task. A systematic review of driver drowsiness detection systems has highlighted that many existing approaches rely on datasets that often lack sufficient diversity in cultural and behavioural profiles. This shortfall can introduce biases, reducing model performance in regions where driving behaviours differ from the training data. Researchers argue that addressing these issues requires adaptive learning techniques and the inclusion of cross-cultural behavioural data, ensuring that detection models are robust, equitable and effective in real-world global deployments (Saleem et al., 2023).

9.1.4 Ethical and bias considerations in CAVs

The implementation of AI in connected automated vehicles for drowsiness detection presents significant ethical challenges that require careful consideration. This section examines the key ethical dimensions and proposed solutions based on current research.

One of the most pressing concerns is algorithmic bias in drowsiness detection systems. Empirical evidence demonstrates significant performance disparities across demographic groups, particularly in convolutional neural networks and long short-term memory-based systems (Grzelak & Brandao, 2021; Ngxande et al., 2020). These biases primarily affect racial and gender groups, highlighting the need for more inclusive approaches to AI development.

Privacy concerns represent another crucial ethical dimension. Current systems often lack adequate consent mechanisms and may enable intrusive surveillance, particularly through technologies like infrared retina tracking (Jamthe et al., 2022; Krontiris et al., 2020). To address these concerns, innovative technical solutions have been developed, such as the Secure Triplet Loss method for protecting biometric data (Esteves et al., 2021). Furthermore some researchers advocate for limiting data sharing to only life-threatening situations (Oueida et al., 2021). Additionally, there is growing emphasis on the need for robust data protection frameworks and stronger consumer privacy rights in the autonomous vehicle context (Krontiris et al., 2020; Wani et al., 2024).

Moreover, the safety implications of drowsiness detection systems are significant. While some studies report promising accuracy rates - 89% offline and 73% online (Rohlinger et al., 2021) and approximately 90% in other cases (Purohit et al., 2023) - questions remain about the generalizability of these results to diverse real-world populations. The potential for false positives or negatives raises concerns about downstream impacts on driver safety and experience.

The establishment of clear accountability frameworks remains a critical challenge. Recent analytical work emphasizes the importance of transparency, explainability and stakeholder engagement as foundational principles for ethical AI implementation (Krontiris et al., 2020; Wani et al., 2024). However, establishing clear lines of responsibility, particularly in cases of system failures or discriminatory outcomes, continues to be problematic.

9.1.5 Bias and ethical considerations on emotion recognition on CAVs

Emotion recognition technologies have the inference of a person's emotional state as their goal. They achieve this through: a) psychometric scales, for example Tsanas et al., (2016) used the statistical tool of principal component analysis, b) speech, for example Chin et al., (2021), used SVM to obtain a 92,87% accuracy in detecting stability in an out-of-hospital cardiac arrest study, c) facial expression, for example Munsif et al. (2022) achieved a 97% accuracy score through facial image extraction, d) physiological signals, for example Purnamasari et al. (2015) achieved a 92,03% accuracy using a back propagation neural network to extract data from EEG and relative wavelet energy and e) multimodality, for example Hossain (2016) achieved an accuracy of 99,4% by using a Gaussian Mixture Model (GMM) and extracting features from video and audio.

Emotion recognition technologies have many uses, such as within medicine (R.Guo et al. 2024) or in road safety within the context of CAVS (W. Li, 2022), a domain which is the object of this literature review. The key source of bias in this domain is bias across demographic

groups. According to two (US) National Institute of Standards and Technology (NIST) Face Recognition Vendor Tests, higher error rates have been detected in dark-skinned individuals, women and older adults (Grother et al., 2019; Duewer, 2022). Further studies have shown bias in favour of Western, young and posed expressions (Barrett et al., 2019). Bias has been detected both in the datasets and the models used for Facial Expression Recognition (Hosseini et al., 2025).

Some root causes of demographic roots bias are unrepresentative training data bias (Crawford et al., 2019), label noise and cultural mismatch (Ong, 2021), proxy features and spurious correlations and finally evaluation blind spots (Grother et al., 2019).

Besides bias considerations, some serious ethical concerns are connected to the use of emotion recognition technologies in the domain of CAVs. A first notable example revolves around privacy, consent and data minimisation (Chowdhary et al., 2025). A significant amount of sensitive biometric and behavioural data is collected by the in-vehicle emotion monitoring systems. This continuous collection of sensitive data risks infringing privacy if performed without robust consent, appropriate retention limits, and strong anonymisation. A related worry is that this data may be used for purposes other than the original ones, such as law-enforcement or targeted advertising, if the governance of their usage is not strict. A second ethical concern arises from possible harms caused by incorrect emotion inferences. Such harms include unjustified safety interventions, withdrawal of services, the mislabelling of users as “aggressive” or “untrustworthy”. Furthermore, emotional profiling may result in stigmatizing vulnerable groups or be used to surveil or penalize drivers or passengers (Doerfler & Stark, 2024). Finally, a third ethical concern comes to the fore due to the complexity of providing consent in shared and public contexts. Informed consent becomes operationally impractical and difficult in complex situations when the passengers share short rides, use public transport or taxis or find themselves in situations other than simple situations when a car always has the same driver and passengers (Doerfler & Stark, 2024).

Despite the bias and ethical considerations, emotion recognition technologies in CAVS still seem like a useful technology, especially as they have the potential to save lives. Mitigation strategies may come from technical, data centric fixes such as diversifying the collection of training data or conducting multimodal fusion and contextual modelling (Hosseini et al. 2025); they may come from privacy-preserving architectures such as through differential privacy or default privacy-by-design (Chowdhary et al., 2025); they may come through co-design, especially with the participation of vulnerable groups in a transparent way (W. Li, 2024); and finally, they may come through policy and regulatory measures, such as the classification of emotion data as sensitive, mandate impact assessments and restrict non-safety uses (Doerfler & Stark, 2024).

9.1.6 Transparency and explainability

Machine learning using inputs like camera images, Electroencephalogram (EEG) / Electrooculogram (EOG) signals, or vehicle data, is now widely used for in-vehicle drowsiness detection. However, many machine learning models are black boxes whose inner logic is opaque. This opacity poses challenges for safety and user trust. Studies emphasize that *explainability* and *robust validation* are essential to make such systems trustworthy (Hasan et al., 2024).

Research has begun to address these issues. For example, Hasan et al. (2024) note that most physiological-signal based drowsiness detectors lack explainability, focusing instead on accuracy. They show that using techniques like SHAP and partial dependence plots can “unbox the black box”, revealing which features drive the model’s decisions. In their multimodal system (EEG, EOG, ECG signals), leave-one-driver-out validation (a robust, subject-independent test) achieved ~80% accuracy, while the explainability analysis identified clear decision boundaries on the most important physiological features.

Similarly, Garcia-Alcaide et al. (2025) integrated XAI into deep learning for in-cabin monitoring. They found that “*inherent opacity often limits our understanding*” of a drowsiness detector’s outputs, so they use XAI to debug models and give transparency on why alerts are triggered. In short, recent studies (Garcia-Alcaide et al., 2025; Hasan et al., 2024) demonstrate how XAI methods (e.g. SHAP, saliency maps, LIME) and rigorous cross-subject validation can make ML-based drowsiness systems more transparent and reliable.

These academic findings align with industry insights. A 2024 review by the Norwegian Institute of Transport Economics notes that machine learning is now used in driver monitoring and fatigue detection, but faces black box and bias issues that demand explainability (Kielland et al., 2024). The report urges that safety-critical machine learning must yield “*human-interpretable, explainable outputs*”, with documented algorithms and data subject to audit. It also notes that ISO standards (ISO 26262 for functional safety and ISO 21448/SOTIF for safety-of-function) require rigorous testing, validation and fail-safe mechanisms for any in-vehicle system.

9.1.7 Future directions to bridge ethical gaps in CAVs domain

Future directions recommendations span from technical, regulatory and operational domains. This synthesis draws from multiple studies to outline concrete pathways forward.

A primary set of solutions focuses on protecting user privacy while maintaining system functionality. On-board data processing emerges as a key technical solution, allowing vehicles to process sensitive information locally rather than transmitting it to external servers (Krontiris et al., 2020). Data minimization principles, including collecting only essential information and implementing automatic deletion protocols, form another crucial component (Fossa et al., 2022). The research advocates for layered transparency approaches, where users have clear visibility into what data is collected and how it’s used, combined with harmonized legal frameworks to ensure consistent privacy protection across jurisdictions (Santoni de Sio, 2021).

To address the black box nature of automated vehicle decisions, researchers propose implementing explainable artificial intelligence systems that can provide clear justifications for their actions (Sütfeld et al., 2019). This includes developing user-centered explainability methods that present information in accessible formats for different stakeholder groups. Algorithm auditing frameworks are recommended to ensure ethical compliance and detect potential biases before deployment (Thornton et al., 2017). These solutions are considered achievable in the short to medium term, as many necessary tools are already in development.

The development of comprehensive safety standards represents a medium to long-term solution requiring industry-wide cooperation. Researchers propose creating surrogate safety metrics that can evaluate vehicle performance across different scenarios (Papadimitriou et al., 2022). These metrics would be complemented by harmonized safety benchmarks and interdisciplinary frameworks that integrate ethical considerations into safety assessments. The

solution framework includes standardized testing protocols and clear performance thresholds that manufacturers must meet.

To clarify liability and responsibility issues, studies recommend implementing the Meaningful Human Control framework (Calvert et al., 2020). This approach defines clear chains of responsibility and establishes accountability mechanisms for different scenarios. The solution includes developing specific liability guidelines for manufacturers, operators and users, supported by technical systems that can accurately record and report decision-making processes (Dogan et al., 2020).

9.2 Accessibility checker for blind visitors of website

Web accessibility ensures that websites, tools and digital content are designed and developed so that people with disabilities can perceive, understand, navigate and interact with them effectively. It encompasses a wide range of considerations, including visual, auditory, motor, cognitive and neurological capacities. As digital platforms increasingly structure access to education, employment, health information and public services, accessibility becomes a critical factor in digital inclusion. However, many users are still excluded from full participation online. A blind user relying on a screen reader may encounter unlabelled buttons or images that lack alternative text, making it impossible to interpret menus or access essential information (Borodin et al., 2010). A deaf user may find a video explaining how to escape during a climate emergency but be unable to follow it because captions are missing (Calgaro, 2024). Someone with a motor impairment may struggle to complete an online form if it requires precise mouse use and does not allow keyboard navigation (Ribera et al., 2015).

These forms of exclusion are well documented across the European Union. In Romania, an automated audit of 60 municipal websites revealed that none conformed to the basic Level A requirements of the Web Content Accessibility Guidelines (WCAG) 2.0 standard⁴⁰ (Pribeanu et al., 2012). The study reported a total of 4,146 accessibility errors on homepages alone, with an average of 69 errors per site. The most common issues were the absence of alternative text for non-text content and the use of HTML tags for visual presentation instead of CSS, both of which severely hinder screen reader functionality. These two issues accounted for over half of the total errors and disproportionately affected blind and visually impaired users. Further problems included redundant link text, poor heading structure and missing labels for form controls. Importantly, the study noted a lack of sustained compliance over time, with many websites deteriorating in accessibility after redesigns, suggesting that accessibility checks were not part of routine release cycles. This pattern is echoed in other EU countries. For example, in Italy, Valtolina & Fratus (2022) reviewed 7,713 municipal portals and found that only 12 percent of homepages were free from accessibility violations. Recurrent issues included insufficient text contrast, unstructured navigation and missing descriptions for visual elements. In Poland, Król & Zdonek (2020) evaluated 182 municipal websites in the Małopolskie Voivodeship using automated tests and a cognitive walkthrough. They reported that the websites achieved only 57,23 percent of the possible points in their Aggregate Accessibility Rating (AAR), with many sites lacking features critical for visually impaired users, such as meaningful structure, keyboard navigation and screen reader compatibility. In Cyprus, empirical research confirms the persistent inaccessibility of many public digital platforms. Iseri

⁴⁰ <https://www.w3.org/TR/WCAG21/>

et al. (2023) evaluated the websites of 38 higher education institutions on the island using WCAG 2.0 criteria. The findings revealed that none of the websites were free of accessibility errors and the majority failed to reach even the minimum acceptable level of compliance. Common barriers included missing alternative text for images, poor colour contrast, lack of keyboard navigation and improperly labelled form fields.

In light of these persistent issues, the WCAG have become the key reference framework for ensuring digital accessibility. Developed by the World Wide Web Consortium (W3C), WCAG outlines concrete, testable success criteria aimed at making digital content accessible to a broad range of users. The guidelines are built around four foundational principles: content must be Perceivable, Operable, Understandable and Robust. These principles are not theoretical but highly practical. For instance, Perceivable content includes captions for multimedia and alternative text for images, while Operable interfaces require full keyboard functionality. The Understandable principle emphasises consistency and clarity in language and structure and Robust content must remain compatible with assistive technologies as they evolve. WCAG has been explicitly integrated into the EU's legal framework through the Web Accessibility Directive (Directive (EU) 2016/2102)⁴¹, making it a central tool not only for technical compliance but also for upholding the rights of disabled users to access public digital services. By providing a standardised, widely adopted structure, WCAG enables public institutions to identify and eliminate barriers that would otherwise prevent equal access. It is not simply a technical checklist but a normative benchmark for digital inclusion, guiding the development of more equitable and user-friendly services across the EU.

Yet despite the clarity of these guidelines, inaccessible design remains widespread. This is not simply a matter of oversight or limited technical knowledge, it reflects a deeper ableist bias in how websites are imagined and developed, often with an implicit assumption of a "standard" user who is sighted, hearing, cognitively typical and uses a mouse and screen in standard ways. In response to this, we developed the ALFIE use case to actively challenge these design assumptions through the development of an AI website checker for blind users.

9.2.1 State of the art for accessibility and AI

Artificial Intelligence (AI) technologies are increasingly influencing the design of tools and systems intended to improve accessibility for individuals with disabilities. AI encompasses machine learning, computer vision, natural language processing and sensor-based modelling, among other techniques. These are often embedded in mobile applications, wearable devices and adaptive software to support independent living, education and communication. To give some examples, Bhagat et al. (2023) conducted a comparative evaluation of four leading applications, Seeing AI, Supersense, Lookout and Envision, reporting gains in accuracy and independence but noting usability limitations, especially when contextual or cultural variables were not accounted for. In higher education, students using AI-supported learning tools demonstrated better access to course materials and greater academic engagement, though the effectiveness varied with platform design and institutional support (Bhagat et al., 2023).

Automatic speech recognition (ASR) is a central AI technology used to support individuals who are deaf or hard of hearing. Recent studies have evaluated how large language models can reduce word error rates and improve real-time captioning accuracy in educational and

⁴¹<https://digital-strategy.ec.europa.eu/en/policies/web-accessibility-directive-standards-and-harmonisation>

professional settings. Fathallah et al. (2024) found that incorporating LLMs into captioning systems resulted in significantly better performance compared to traditional ASR tools. However, Kuhn et al. (2024) reported persistent challenges when dealing with background noise, technical vocabulary and diverse speech patterns. These findings suggest that while AI is improving communication access, robust context adaptation and human oversight remain necessary for reliability.

AI is also contributing to interaction support for people with limited mobility through gesture recognition, voice interfaces and adaptive input systems. While the literature in this area is less extensive, studies have found that AI-driven assistive technologies, such as head movement-based controls and voice-activated systems, improve task performance and user satisfaction (Korada et al., 2024). In comparative testing, AI-enhanced tools were shown to reduce task time and increase perceived control among users with motor impairments. However, challenges remain in calibrating these systems to individual movement patterns and environmental variability.

AI-based systems have shown promise in supporting individuals with cognitive disabilities and neurodevelopmental conditions such as autism, ADHD and intellectual disability. Perry et al. (2024) conducted a comprehensive review and concluded that AI-enabled tools could enhance adaptive functioning, communication and social interaction. In educational contexts, Yang and Taele (2025) introduced *Audemy*, an audio-based adaptive learning platform tailored for blind and neurodiverse learners. Their study showed increased user engagement and knowledge retention, especially when teachers were involved in adapting the content. These studies emphasize the importance of personalization, accessibility-by-design and trust in AI deployment for cognitive support.

AI is being increasingly applied in inclusive education, particularly for content adaptation, reading support and assessment tools. Gibson (2024) reported that AI can automate the generation of accessible learning materials, including alt-text, audio narration and layout adaptation. Such systems have been shown to benefit not only students with documented disabilities but also a broader population of learners with varied needs. Nonetheless, ethical questions remain around data collection, bias in learning models and the risk of over-dependence on AI systems in pedagogical contexts.

9.2.2 Challenges of accessible AI and bias

Recent years have seen a marked increase in research exploring the role of artificial intelligence in improving digital accessibility. A systematic review by Chemnad and Othman (2024) analysed 3,706 articles and narrowed them down to 43 relevant studies that applied AI methods to address accessibility challenges. This work revealed that most AI applications have concentrated on support for users with visual impairments, particularly in generating image descriptions and enhancing screen reader usability. Other disability domains, such as cognitive, auditory and motor impairments, remain significantly underrepresented in the current body of research.

9.2.3 Ethical considerations

Ethical concerns have been raised regarding data bias, transparency and exclusion in AI-based accessibility tools. Morris (2019) proposed a framework for evaluating AI applications through the lens of disability justice, emphasizing that assistive AI must not only function reliably but also respect user agency, privacy and social context. Bibliometric reviews have



revealed an uneven research distribution, with a heavy emphasis on visual accessibility and relatively little work on auditory, cognitive and motor domains (Naayini et al., 2025). Many systems are also found to fall short of accessibility standards, suggesting a gap between technical potential and actual inclusive implementation.

9.2.4 Future directions

Accessible AI is a double-edged sword: it democratizes power, but it also democratizes risk. As AI becomes more pervasive, bias will become less a technical flaw and more a societal failure, hence the timely development of ALFIE. On the one hand, as AI tools become more accessible (through low-code platforms, APIs, etc.), they'll be adopted by non-experts. These users may not understand how bias can creep in or how to mitigate it. Users also may depend on unreliable data to build models, which in turn will be biased, unreliable and lead to inequalities reflecting cultural, racial, gender and socioeconomic biases. On the other hand, when training AI, we need to prioritise regulatory and ethical awareness. It is equally important to address contextual and cultural issues in AI training. AI training, data and tool availability will lead to AI democratisation, which means that more sectors will use AI, but they might lack guidelines to spot and handle bias, again, one of ALFIE objectives in the development of the ETD-Hub and AutoML. Without strong domain-specific ethical checks, AI use could become opaque and unaccountable in sensitive environments such as health.

AI outputs are used as data for future models; bias can compound over time in actions such as predictive policing and hiring algorithms. This creates a self-fulfilling loop of inequality, increasingly hard to correct without major intervention. "Accessible AI" often still means access for wealthier or more digitally connected regions. Developing nations, vulnerable groups and minority languages may lack localized data, infrastructure, or cultural adaptation in AI tools, leading to biased and irrelevant applications. Finally, open access to powerful AI tools may encourage bad actors to exploit bias intentionally, for example for disinformation, targeted harassment. Democratization without responsible design increases the threat of weaponized AI bias, with deliberate or negligent use of biased AI systems to harm individuals, communities, or public trust. When AI becomes widely accessible but is not designed responsibly, it becomes easier for bad actors to exploit its weaknesses and for well-meaning users to cause harm unknowingly.

9.3 Compliance screening for partners

The integration of AI into compliance screening for business partners represents both tremendous opportunity and significant ethical complexity. Organizations across financial services, supply chain management and other regulated industries are deploying sophisticated AI systems to enhance efficiency while navigating increasingly complex regulatory requirements. This section examines the current state of compliance AI, identifies critical ethical challenges and outlines emerging solutions for responsible implementation.

9.3.1 State-of-the-art algorithms for compliance monitoring

Modern compliance screening systems utilize advanced machine learning techniques that offer significant operational improvements across various domains. Cutting-edge algorithms have revolutionized traditional rule-based compliance methods by incorporating adaptive learning, improved pattern recognition and real-time risk assessment capabilities that greatly surpass those of conventional systems.

The core architecture of contemporary compliance monitoring systems relies on ensemble methods that combine multiple algorithmic approaches to achieve robust performance. Support Vector Machines (SVMs) have proven particularly effective in compliance classification tasks due to their ability to handle high-dimensional data and create optimal decision boundaries in complex feature spaces (E. Mienye et al., 2024). Kalusivalingam et al. (2022) examined the implementation of this technique to streamline governance processes, ensure regulatory adherence and mitigate risks associated with non-compliance. However, they found that algorithmic bias presents a critical limitation.

Random Forest algorithms have also demonstrated themselves to be suitable solutions for compliance risk assessment due to their ensemble nature and robustness to overfitting. In compliance applications, these algorithms excel at handling the mixed data types commonly found in business partner screening, including categorical regulatory flags, numerical financial metrics and text-based risk indicators. For example, Stewart (2025) shows that this kind of algorithm can analyse large datasets to identify irregularities, outliers, or patterns that deviate from established norms. These anomalies can indicate potential compliance violations, fraud, or risks that would be difficult to detect through manual reviews.

Multi-layer perceptrons with backpropagation training have shown particular effectiveness in processing textual compliance documents and extracting risk signals from business communications (Mienye et al., 2024). Dimlioglu et al., (2023) proposed an automatic document classification pipeline using this technique that determines whether a document is important for the company or not and if deemed important, it forwards those documents to the departments within the company for further review. Recent developments in deep learning have also revolutionized compliance monitoring capabilities, particularly in natural language processing applications. Transformer architectures have demonstrated exceptional performance in processing regulatory documents and extracting compliance-relevant information from multilingual business correspondence (Bednár et al., 2025). The attention mechanism enables these models to focus on relevant regulatory concepts while maintaining context across long documents.

9.3.2 Challenges and fairness in compliance systems

The regulatory landscape for compliance AI involves overlapping frameworks that organizations must navigate simultaneously. The GDPR Article 22 prohibits decisions based solely on automated processing that produce legal effects, requiring meaningful human oversight for partner screening decisions (Wachter et al., 2018). This creates tensions between automation efficiency and regulatory compliance requirements.

The European Union's AI Act introduces risk-based classification frameworks that impose stringent requirements on high-risk AI systems, including many compliance tools. Organizations must conduct comprehensive risk assessments, ensure high-quality datasets with bias minimization, maintain detailed technical documentation and implement conformity assessments (Veale & Borgesius, 2021). Systems affecting credit scoring, employment decisions, or essential service access qualify as high-risk, requiring extensive pre-market and operational compliance measures.

Data quality challenges compound compliance complexity through multiple pathways. Historical training data often exhibits systematic underrepresentation of certain business types, temporal bias from changing regulatory environments and selection bias favouring larger



entities with extensive documentation histories (Barocas & Selbst, 2016). These data quality issues create systematic gaps that disadvantage specific partner categories and may perpetuate existing market inequalities.

Geographic disparities arise from differences in infrastructure and regulatory maturity across jurisdictions for large companies. Compliance systems may inadvertently reinforce existing market inequalities through proxy discrimination via factors such as company size, industry sector, or geographical location (Kalluri, 2020). Small and medium-sized enterprises face particular disadvantages due to limited historical data availability and resource constraints that affect the quality of compliance documentation. Additionally, Kleinberg et al. (2016) proved that demographic parity, equal opportunity and calibrated fairness cannot be simultaneously satisfied except in trivial cases. This forces organizations to make explicit trade-offs between different fairness measures based on regulatory requirements and stakeholder values.

9.3.3 Ethical and bias considerations in compliance screening

Compliance AI systems demonstrate systematic biases that disproportionately impact certain business categories and geographic regions. Research reveals that algorithmic systems exhibit cultural bias when deployed across diverse jurisdictions without proper localization (Blodgett et al., 2020). Business practices considered normal in some cultures may be incorrectly flagged as irregular or risky by algorithms trained primarily on Western business data.

Algorithmic bias manifests through multiple pathways in compliance screening systems. Historical training data perpetuates existing discriminatory practices, as demonstrated by research on criminal justice algorithms like COMPAS (Mattu, 2016). These studies show how biased datasets can amplify discriminatory outcomes in automated decision-making systems, with implications extending to business partner screening applications.

Privacy concerns represent another critical ethical dimension in the compliance of AI systems. The collection and processing of extensive business relationship data raise fundamental questions about data minimization, purpose limitation and consent mechanisms (Solove, 2006). Organizations must balance comprehensive risk assessment capabilities with privacy protection requirements, particularly when processing sensitive information about business partners' financial status, geographical locations, or industry associations.

The challenge of proxy discrimination emerges when seemingly neutral variables correlate with protected characteristics. Business size, transaction patterns or geographical location may serve as proxies for characteristics that should not influence compliance decisions. Barocas & Selbst (2016) demonstrate that algorithmic systems can perpetuate discrimination even when explicitly protected attributes are removed from training data, as correlated variables maintain discriminatory signals.

9.3.4 Transparency and explainability in compliance AI

Regulatory demands for explainable AI have accelerated the development of sophisticated interpretation techniques. The EU AI Act mandates comprehensive transparency for high-risk systems, while GDPR's "right to explanation" requires individual decision justification (Goodman & Flaxman, 2017). Post-hoc explanation methods like LIME and SHAP have become standard approaches for interpreting complex compliance models (Ribeiro et al., 2016).

Recent advances in explainability techniques focus on audience-specific explanation frameworks that tailor technical details to the needs of different stakeholders. Compliance officers require a different level of explanation granularity compared to business partners or regulatory auditors. As a result, the effectiveness of explanations heavily depends on the recipient's technical background, domain expertise and specific information needs (Miller, 2019).

The challenge of explanation stability arises when similar compliance cases receive different explanations due to model complexity or feature interactions (Alvarez-Melis & Jaakkola, 2018). Explanation consistency across comparable business scenarios is essential for regulatory acceptance and stakeholder trust, prompting the development of more stable interpretation methods that offer coherent justifications for similar compliance decisions.

9.3.5 Future directions to bridge ethical gaps in compliance domain

Emerging standards provide comprehensive frameworks for ethical compliance AI implementation. ISO/IEC 42001:2023⁴² represents the first comprehensive AI management system standard, offering risk-based assessment throughout the AI lifecycle with specific fairness integration requirements.

Technical innovations in explainable AI focus on audience-specific explanation frameworks, real-time interpretation capabilities and hybrid approaches balancing performance with transparency. Advanced bias mitigation combines pre-processing data transformation, in-processing fairness-aware algorithms and post-processing output adjustments with continuous monitoring systems providing automated bias detection and response protocols (Friedler et al., 2019).

Organizations are increasingly establishing AI ethics boards with cross-functional expertise and integrated governance programs that combine organizational and technical controls. Enhanced human-AI collaboration models emphasize AI-assisted human review with mandatory oversight for final compliance decisions, clear escalation protocols for complex cases and continuous learning loops where expert feedback improves system performance (Amershi et al., 2019).

Privacy-enhancing technologies progress through improved differential privacy mechanisms, innovative federated learning architectures for regulatory applications and homomorphic encryption for sensitive compliance computations. These advancements facilitate collaborative compliance intelligence while maintaining data protection requirements across organizational boundaries (Kim et al., 2025).

Research directions emphasize the development of culturally aware compliance models that account for regional variations in business practices, the implementation of robust bias detection frameworks that monitor intersectional discrimination patterns and the creation of standardized fairness metrics specifically tailored to compliance screening contexts (Mehrabi et al., 2021). Future systems will likely incorporate automated bias correction mechanisms, enhanced explanation stability methods and cross-jurisdictional regulatory compliance frameworks that adapt to evolving legal requirements.

⁴² <https://www.iso.org/standard/42001>

The convergence of emerging international standards, advancing explainable AI techniques and evolving regulatory frameworks presents opportunities for the responsible deployment of AI. Organizations that proactively adopt ethical frameworks, invest in bias mitigation strategies and prepare for regulatory evolution will be best positioned for sustainable success in compliance screening applications.

9.4 Identified ethical and bias concerns

The current literature review highlights the multifaceted nature of modern AI deployments and the significant ethical gaps that accompany their increasing integration across various sectors. From CAVs to web accessibility and compliance screening, AI's transformative potential is undeniable, yet its responsible development and deployment are hampered by persistent challenges related to bias, transparency, privacy and accountability.

A recurring theme across all examined domains is the pervasive issue of algorithmic bias. In CAVs, models designed for driver drowsiness detection exhibit performance disparities across demographic groups, suggesting a "one-size-fits-all" approach overlooks crucial physiological and behavioural variations. This bias is often rooted in unrepresentative training data, leading to unequal safety outcomes in real-world scenarios. Similarly, in web accessibility, existing AI applications disproportionately focus on visual impairments, leaving other disability domains significantly underrepresented and reinforcing existing inequalities. The implicit assumption of a "standard" user in website development further exacerbates this issue, reflecting a deeper ableist bias. In compliance screening, historical training data can perpetuate discriminatory practices, creating systematic gaps that disadvantage specific business types or geographical locations. Algorithmic systems can even exhibit cultural bias when deployed across diverse jurisdictions without proper localization, incorrectly flagging normal business practices as irregular. The difficulty in simultaneously satisfying demographic parity, equal opportunity and calibrated fairness further complicates efforts to create unbiased compliance systems.

Transparency and explainability emerge as critical ethical challenges, particularly due to the black box nature of many AI models. In CAVs, the opacity of machine learning models for drowsiness detection poses significant safety and trust concerns. While XAI techniques are being developed to reveal model decisions, the need for rigorous validation and human-interpretable outputs is paramount. For web accessibility, the lack of transparency in AI tools can hinder understanding of how content is adapted or why certain accessibility features are missing. In compliance screening, the "right to explanation" mandated by regulations like GDPR and the EU AI Act highlights the demand for clear justifications for automated decisions, a challenge compounded by model complexity and feature interactions.

Privacy is another prominent ethical consideration. In CAVs, systems often lack adequate consent mechanisms and can enable intrusive surveillance, particularly through technologies like infrared retina tracking. The balance between comprehensive risk assessment capabilities and privacy protection is also a critical issue in compliance screening, especially when processing sensitive business partner data. The potential for extraction of training data from large language models further underscores the need for robust privacy-enhancing technologies.

Accountability frameworks remain a significant challenge in all AI deployments. Establishing clear lines of responsibility in cases of system failures or discriminatory outcomes is problematic, especially with the intricate decision-making processes of AI. The lack of

comprehensive ethical guidelines and clear legal frameworks for AI development and deployment contributes to this ambiguity.

9.5 Future directions to bridge ethical gaps

Addressing the ethical gaps, a multi-pronged approach encompassing technical, regulatory and operational domains is required.

From a technical perspective, advancements in privacy-enhancing technologies are crucial. On-board data processing, data minimization principles and automatic deletion protocols can protect user privacy in CAVs. For compliance, improved differential privacy mechanisms, federated learning architectures and homomorphic encryption can facilitate collaborative intelligence while maintaining data protection. The development of robust bias mitigation strategies, including pre-processing data transformation, fairness-aware algorithms and continuous monitoring systems with automated bias detection and response protocols, is essential across all domains. Furthermore, technical innovations in explainable AI, such as audience-specific explanation frameworks and real-time interpretation capabilities, can enhance transparency and trust.

Regulatory frameworks are vital in shaping responsible AI development. The EU AI Act, with its risk-based classification and stringent requirements for high-risk systems, provides a foundational legal framework. However, its effective implementation depends on coherent and equitable application across the European Union, with a focus on clear definitions, guidelines and enforcement mechanisms. The development of harmonized standards, such as ISO/IEC 42001:2023, provides comprehensive frameworks for ethical AI management systems with specific fairness integration requirements. The establishment of AI ethics boards with cross-functional expertise and integrated governance programs will be critical for proactive ethical oversight.

Operational strategies emphasize enhanced human-AI collaboration. This involves AI-assisted human review with mandatory oversight for final compliance decisions, clear escalation protocols for complex cases and continuous learning loops where expert feedback improves system performance. For CAVs, this translates to ensuring drivers remain vigilant and ready for manual takeover in semi-automated systems, with drowsiness detection serving as a critical safety mechanism. In web accessibility, a shift from an ableist bias in design to an "accessibility-by-design" approach, where accessibility is integrated from the outset, is crucial. Training data for AI models needs to be more diverse and representative, accounting for cultural and behavioural differences to prevent biased outcomes.

Ultimately, bridging the ethical gaps in modern AI deployments necessitates a proactive and collaborative effort from developers, policymakers, researchers and users. The goal is to democratize the power of AI while simultaneously mitigating its risks, ensuring that its benefits are realized equitably across all segments of society without perpetuating existing inequalities or creating new forms of harm.

10 Ethical considerations

10.1 Current AI legislation

The European Parliament passed the AI Act in 2024. It is the first comprehensive legislation about the regulation of AI worldwide (Regulation (EU) 2024/1689). The Act classifies AI according to an assessment of risk to its citizens. There is a substantial focus on preventing the use of high-risk applications such as the social scoring of individuals, use of biometrics and manipulation of information (EU AI Act⁴³). The Act has some provision for lower risks. In addition, there is current legislation and regulation that will apply to AI and its developments and will continue to evolve to adjust to social and economic change mitigated by AI. The GDPR will remain important in this respect. Also, the development of standards by International and European Standardization Development Organisations (SDOs) like the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC), European Telecommunications Standards Institute (ETSI), Institute of Electrical and Electronic Engineers (IEEE) and the International Technology Union Telecommunication Standardization Sector (ITU-T) as part of the United Nations (Righi et al, 2022: pages 41-43).

Article 3 of the Act gives fundamental definitions. There is a specific definition of what Artificial Intelligence is. Key inclusions are ‘machine based’ learning, with a level of ‘autonomy’ and the ability to ‘adapt’ via learning (Article 3: Definitions | EU Artificial Intelligence Act⁴⁴). A ‘provider’ is a person or organisation who develops an AI system. A ‘deployer’ is an organisation using AI (not including individual use). An ‘operator’ is an all-embracing term that includes both ‘providers’ and ‘deployers’ and those other organisations connected in some way to AI.

Before moving to a review of the current literature about the challenges of implementing the Act and evolving a satisfactory system of regulation, this next section reviews the key historical literature on the wider context of ‘public policy’. This is a contemporary view of continental policy systems and their complex interaction with national governments and a wider range of policy stakeholders.

Contemporary policy theory focuses on the importance of ‘governance’ as a wide-ranging dynamic policy process with numerous stakeholders and potential determinants. This progression from an orthodox focus on national political ‘government’ as the primary driver of public policy has become increasingly important in recent decades. This is a seminal development in theorising public policy, the turn from ‘government to governance’ (Rhodes, 1997).

A governance perspective on the dynamic relationship between AI and government intervention argues that there are many factors to consider and that it is difficult for a democratic government process to ‘choose’ a single ethical policy path or to predict what will cause ‘good’ outcomes over ‘bad’ outcomes. Given that the economic and social trajectory and impact of AI is unclear, one argument about the development of policy is that it should focus on preventing misuse (Büthe, et al, 2022).

⁴³ <https://artificialintelligenceact.eu/high-level-summary/>

⁴⁴ <https://artificialintelligenceact.eu/article/3/>

All policies designed and legislated for by a political government have to be implemented by complex aspects of the wider governance process, if they are to be successful in delivering what the political designers want to achieve. This implementation complexity includes the relationship between national governments with international and continental governance organisations and regulators, regional and local governments, transnational and national business interests and non-government organisations (NGOs), including professional bodies and trades unions.

If policy is to be successful in such a dynamic and unstable world, politicians and some stakeholders need to develop a coherent intervention and regulation strategy based on trust, negotiation and mutual adjustment. For some commentators, this relational approach is best understood as ‘policy networks’ (Klijn & Koopenjan, 2000). Agreeing a set of primary and shared values that highlights a common purpose becomes a central part of this networked and relational approach to steering governance (Haynes, 2015). Rhodes (2000) implies that the move from ‘government to governance’ leads to marketisation and contracting public services to private companies and results in a loss of trust in the public policy process. Meadows (2009), the systems theorist and environmentalist, suggests that nurturing core and shared beliefs about what is most important is pivotal to success in achieving optimal social outcomes. Agreeing an ethics for the design and application of AI is therefore a vital component in the current policy process and to maximise the public benefits of AI. The European Commission had a specific project before the writing of the AI Act legislation to try and achieve such prior ethical agreement. This resulted in the *Ethics guidelines for trustworthy AI*, as discussed earlier in this literature review (European Commission, 2019).

Implementation theory also amplifies the complexity of public policy processes and their inherent unpredictability and indeterminate outcomes (Hill & Varone, 2021). Major historical contributions to implementation theory include the ‘top down’ perspective, with its conceptualisation of the ‘implementation deficit’. This is a list of events and process errors that can prevent legislation achieving its target social and economic outcomes. Two examples (there are many more) are: (i) failures in inter-organisation communications and (ii) an inadequate theorisation of the underlying social problem (that the policy is attempting to change). Also, implementation theory has conceptualised a ‘bottom up’ perspective that focuses on important micro and meso influences that exist after the policy legislative process. These include stakeholders who work close to the social practices defined by policy ideas. They change and undermine policy objectives and this includes actions by professional bodies and organisations awarded government contracts. Here the day to day working of policy into practice is at the ‘front line’ (for example in hospital wards and school classrooms) and involves real world interpretations of rules and standards so that practical interventions are achievable in the daily performance of practice. Local practice is defined by local agencies and professionals. They want policy to make sense for their own local agendas and circumstances, so that outcomes can become rather different to what the national legislature intended. As an example, commentators will see the differences across Europe evident in the practices of the GDPR, as being evidence of differing local practices by Data Protection Officers within organisations and diverse interpretations of the required European standards by the appointed national regulators.

Maggetti (2025:4), however, argues that regulation is the primary mode of contemporary governance practice. Progress depends on the ability of the regulator to determine their priorities and preferences and then take action, without undue influence. Also, the relationship

of the appointed regulator-agent with their appointing political-principal, must be capable of adapting and evolving to face external contingencies and changing complexities (p.10).

In summary, the historical basis of public policy theory and practice dictates that the formation of government intervention and regulation in the new field of AI will be inherently complex and problematic with considerable ‘unknowns’ and challenges. The role of values, standards and their regulation are at the core of this complexity.

10.2 The process of implementation

A recent critical overview of the EU AI Act analyses its functional structures and complexities (Sousa e Silva, 2025). The implementation of the Act depends on organisations and processes at the national and EU level. At the centre of the EU and its Commission are the EU AI Office and EU AI Board, as well as an Advisory Forum and Scientific Panel. The European AI Office will sit within the Directorate-General for Communication Networks, Content and Technology (DG CNECT)⁴⁵. Within individual nation states there exist a notifying authority and a market surveillance authority.

- The European Commission’s AI Office will supervise general purpose AI models and their marketisation. It will have supervisory powers and can impose penalty fines.
- The AI Board has representation from member states and attempts to maintain coherence of standards across the EU.
- In addition, The European Parliament has a Working Group on the Implementation and Enforcement of the AI Act.

Sousa e Silva (2025) warns that the complexity of the legislation makes implementation difficult and can discourage economic investment in AI when compared to other countries outside of the EU. It is, however, acknowledged that if the adoption of EU wide AI standards of practice are coherent and equitable across the Union, compliance costs will be low, resulting in a stable, competitive market with confidence in production and delivery established. The details and relational aspects of implementation are therefore vital to the success of governance and regulation.

The European Commission’s ability to implement and regulate depends on established committee and consultation procedures. *Implementing acts* apply EU laws consistently across member states, following collaboration with a committee of member state representatives. *Delegated acts* modify operational details of European legislation and need member state consultation but not via a formal committee procedure. The European Parliament and Council have two months to raise any objections. Then the delegation proceeds, normally for five years and while subject to scrutiny from the Parliament and Council (Novelli, et al, 2025: p5).

The AI Act implementing agencies (i.e., the AI Office and AI Board) will develop relevant implementing and delegated acts and manage the latter. Their activities will include the following aspects (as summarised by Novelli, et al, 2025, p.6, table 1): working with the European policy implementation processes, issuing guidelines, classifying risk, regulating prohibited systems, harmonising technical standards and information obligations and overseeing enforcement.

⁴⁵https://commission.europa.eu/about/departments-and-executive-agencies/communications-networks-content-and-technology_en

10.3 Risk

Much of the EU AI Act is focused on the assessment and management of risks from AI systems, products and their use. The implementation agencies will author and circulate guidelines about the definitions of AI systems specified in the AI Act. This includes the prohibition of unacceptable and high-risk systems and practices (Article 5). They will develop risk assessment methods and consider rules for modifications of the risk level of high-risk systems (Article 9). This includes the use of conformity assessments for high risks. Conversely, there will be criteria for exceptions to prohibitions of high risk, for example to facilitate law enforcement using biometric identification. On 4th February 2025, the EU Commission published its guidelines on prohibited AI practices⁴⁶.

Implementation agencies will use delegated acts to update the addition and removal of AI systems defined as high-risk (Article 7) and to modify the classification of General Purpose Artificial Intelligence (GPAI) when it demonstrates a “systemic risk”.

High-risk systems must satisfy ‘essential requirements’ under Articles 8 - 15 of the Act, Chapter III, Section 2 (Smuha & Yeung, 2025, p. 238). These include:

- Comprehensive use of a risk management system (Article 9).
- Data management that trains, validates and tests to ensure correct representation and use, including the monitoring for biases (Article 10).
- Adequate technical documentation and automatic logging of the system (Articles 11-12).
- Transparency of the systems workings to enable deployers insights into purpose, capabilities and limitations (Article 13).
- Overseeing of the system by human actors (Article 14).
- Robust accuracy and cybersecurity (Article 15).

Health care AI systems and applications are highly likely to be considered as high risk (Article 6). For this reason, health professional organisations, like the European Society for Radiology (ESR) are keen to be involved early in the formulation of the implementation plans for the Act and wish to work in partnership with the specialist European Commission AI agencies (Kotter, D’Antonoli, Cuocolo, et al. 2025).

Other high risk AI system domains mentioned in Annex III of the Act include: biometric identification, emotional recognition, critical infrastructure, education and training, employment and human resources, provision of personal public and private financial and insurance services, law enforcement, migration and border control and the administration of government services (Smuha & Yeung, 2025, pages 239-240).

AI is considered to be a potential risk when a system design indirectly creates the possibility of misuse. For example, misuse in the form of impersonation, manipulation or deception (via chatbots and deep fakes). In such circumstances, regulators require information and transparency ‘obligations’ (Article 50). A user must be informed that information is being provided by AI rather than a human providing a personal service and consent must be given for the collection of personal data. Exceptions are permitted for law enforcement and criminal

⁴⁶<https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-prohibited-artificial-intelligence-ai-practices-defined-ai-act>

justice interventions. Likewise, users should be made aware when creative content like audio, images and texts are artificially generated.

10.4 Technical documentation and standards

The Commission's specialist agencies will produce requirements for AI technical documentation that will provide a degree of transparency about AI systems. They will approve codes of practice. There will be specific requirements about information transparency for high-risk systems and for GPAI.

Service providers using GPAI have an obligation to write and update technical documentation. This must comply with EU law by having sufficient detail, including evidence of data training before models go to market (Smuha & Yeung, 2025, p243). In some cases, providers of GPAI models can conduct their own audits rather than it being done by an independent party.

On July 10th, 2025, The Commission published its draft, voluntary code of practice for providers of GPAI⁴⁷. It is now being considered by the relevant specialist bodies of the Commission and by Member States. If accepted in its final form, it reduces the regulatory administrative burden for those who voluntarily sign it and provide more legal certainty as they take their product to market.

The code of practice is welcomed by the European Parliament⁴⁸, leading to their implementation and this is resulting in numerous AI providers accepting them as a voluntary code of practice.

Also on July 18th, 2025, the Commission published its guidelines for providers of GPAI⁴⁹. These covers definitions of key concepts related to GPAI and its use and when GPAI is considered 'general-purpose'. Also, the guidelines clarify when GPAI models are considered 'open-source' and how this can permit exceptions from certain obligations.

Zwitter & Gstrein (2024) remark that the approach of the Act to GPAI is important because of its focus on being proactive in risk management and setting appropriate operational standards, rather than being reactive to problems that have occurred.

Kilian, Jack & Ebel (2025) see 'harmonised standards' as the foundation of efficient AI governance under the AI Act. Following a qualitative survey of industry representatives they conclude that implementation needs a substantial lead in time, including adequate consultation with maximum participation from different sectors. They also report that costs, including technical challenges, are formidable and it is important to avoid barriers to market entry, especially for small and medium sized businesses (SMEs).

10.5 Multi-agency cooperation

The Commission's AI agencies will take a view on multi-agency cooperation and the use of existing EU legislative frameworks, such as the GDPR, regarding their interaction with the AI Act. Similarly, they will oversee the activities of member states specialist AI agencies and advise on enforcement and the setting of penalties. They will regulate the AI Act developmental

⁴⁷ <https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai>

⁴⁸ <https://www.europarl.europa.eu/news/en/press-room/20250716IPR29706/lead-meps-react-to-general-purpose-ai-code-of-practice>

⁴⁹ <https://digital-strategy.ec.europa.eu/en/policies/guidelines-gpai-providers>

‘sandboxes’, that are supposed to assist innovation and SMEs, although specific sandboxes are likely to be overseen by the related national authority (Smuha & Yeung, 2025, pages 245-246). Buocz et al (2023) raised concerns during the legislative process of the AI Act about a potential lack of clarity between European Commission versus national government regulatory guidance for the development of sandboxes.

10.6 Critiques of the EU AI Act

Smuha and Yeung (2025) question whether the Act’s regulatory practices will be able to create sufficient oversight of the use of AI and its ability to abuse the human rights and privacies of European Citizens. Their primary criticism is that the legislation delegates too many regulatory tasks to AI providers without sufficient oversight. The authors comment (p. 241) the imposition of new obligations is a positive development, their likely effectiveness is a matter of substantial concern’.

They imply that the philosophy of the Act is more influenced by an economic need to create adequate competitive opportunities for European businesses and services, rather than having a primary purpose to protect human and consumer rights, although they acknowledge the legislation attempts to cover both aspects. This debate is set in a recent historical context of European economic outputs in AI and information services having fallen behind the US and China, causing a dependence on products and services external to the EU (Righi et al., 2022). This is despite evidence of high-quality European activity in research and development.

Kilian, Jack & Ebel (2025) are concerned that the Act’s approach to implementing technical standardisation is too ambitious and this will lead to dysfunctional outcomes. They recommend (p. 32): reducing the number of standards, adjusting implementation deadlines to allow more time, include more support and subsidise start-up costs for SMEs, increase technical expertise in AI advisory groups and committees, provide implementation toolkits and evaluation frameworks and align international and European standards where possible.

Pham & Davies (2025) argue that the AI Act problematises AI, as defined by an exceptionalist notion of Europe as a coherent policy actor and political community and this requires an attempt to define a shared ‘techno-solutionism’. They acknowledge that the Act is attempting to balance the economic opportunities of AI with a threat to the rights of European citizens, but that these policy aims are difficult to reconcile.

Similar to the above, Krarup & Horst (2023) question the ability of the EU AI Act to resolve its desires to promote single market advantages for AI and the free flow of data within the Union while also enabling ethical and responsible AI use for the wider benefits of its citizens. Their study is based on a substantial documentary analysis of official texts underpinning the development of the new legislation.

Tallberg, Lundgren & Geith (2024) research the European AI regulation preferences of business and other non-state actors. They find that all support regulation in some form. Business actors were less concerned about the risks of AI than other non-state actors. These differences between business actors and others were greater in countries with larger commercial AI sectors.

The specific criticisms of the EU AI Act must be viewed in the wider context of the international development of AI in a global marketplace where international relations and trade are increasingly conflicted. Policy analysts argue that the establishment of the governance and

regulation of AI by international, continental, and national organisations is crucial (Kissinger, Schmidt and Huttenlocher, 2024). Some argue that the EU is leading the world in the development of this necessary governance (Dixon, 2024). The US and China, however, are argued to have more volume of investment in, and production of AI related products and services compared to the EU. The role of the US in governance is unclear at the time of writing, primarily due to the change of government in 2025. In recent months, the policy focus has shifted from regulation and the prevention of harm towards de-regulation and maximising innovation and growth.

The AI Index 2024 (Oxford Insights, 2024⁵⁰) for Government AI Readiness includes a country comparison score for Governance and Ethics. This component is a composite from several relevant international indicators and desk research about policy and practice documents. It aims to measure for each country the research question: 'Are there the right regulations and ethical frameworks in place to implement AI in a way that builds trust and legitimacy'? Several EU countries have the highest international rankings for this score. Denmark, Luxembourg, Finland, Netherlands, Norway and Sweden are the top six ranked countries in the world with scores of above 96%. Several other EU countries score above the US (91,14%). These include Germany (94,19%), Ireland (94,14%), Estonia (92,18%), Belgium (92%), Austria (91,72%), and France (91,44%). The United Kingdom also scores above the US with 93,88%. European Union countries that score below the US are Portugal (90,23%), Spain (89,85%), Czechia (89,41%), Italy (89,26%), Slovenia (88,71%), Greece (87,84%), Latvia (87,30%), Malta (86,99%), Slovakia (86,83%), Poland (85,85%), Romania (82,81%) and Hungary (81,41%). All have scores above 81,41%. China scores 68,97% and only three EU countries score below this: Lithuania (68,66%), Croatia (63,88%) and Bulgaria (59,07%). The diversity of scores across the EU evidences the continuing national differences in the Europe Union, despite the continental focus on the EU AI Act and the associated efforts of the European Commission. The future scores for 2025 will provide a better indication of whether the AI governance and ethics practice differences between EU countries are decreasing because of the most recent implementations of the EU AI Act.

While commenting on a draft of the legislation, Veale & Borgesius (2021) argue for more clarity to be provided about the jurisdiction and responsibilities of the Commission in partnership with national governments. Previous research evaluating the use of the GDPR shows evidence of substantial differences about how it is applied by different national states⁵¹. Use of the GDPR, nevertheless, also indicates the types of ethical themes that are most likely to be of primary concern to AI regulators across Europe. For example, CMS law published an extensive analysis of the implementation of GDPR. They report that the most likely areas for enforcement activity are 'insufficient legal basis for data processing', 'non-compliance with general data processing principles' and 'insufficient technical and organisational measures to ensure information security. This provides some evidence about likely areas of difficulty for the providers and deployers of AI in the future.

⁵⁰ *Governance AI Readiness Index 2024*, Malvern: Oxford Insights 2024, <https://oxfordinsights.com/ai-readiness/ai-readiness-index/>

⁵¹ <https://cms.law/en/int/publication/gdpr-enforcement-tracker-report>



10.7 Conclusion

Experience in the last decade with the implementation and application of GDPR shows how important the details of regulatory practice will be for the success of the EU AI Act, in terms of its impact on preventing harmful use of AI without hindering innovation that is in the public interest. Reconciling market advantage to secure economic growth while protecting its citizens will continue to be the fundamental tension at the heart of policy implementation and practice. The Commission will keep a close eye on national differences in the implementation of the Act and its regulations and will seek to ensure there is enough similarity in practices to ensure the equitable operation of the single market and consumer and citizen protections. They will also hope to see European businesses and citizens less dependent on AI services provided by markets outside of the EU.



11 Conclusions

This deliverable has examined the landscape of modern AI deployments through both a technical and ethical perspective, uncovering the growing dissonance between rapid technological progress and the slower pace of ethical, legal and societal alignment. AI has evolved from a niche research field into an omnipresent force shaping sectors as varied as healthcare, transportation, law enforcement and the creative industries. With this shift, AI systems are no longer isolated tools; they are decision-makers, influencers and, increasingly, autonomous actors whose reach extends into deeply human domains.

The analysis has revealed three interlinked patterns. First, AI capabilities are developing faster than the regulatory frameworks, ethical standards and cultural norms meant to guide them. The result is a widening “ethical gap” where technology’s potential and society’s preparedness are misaligned. Second, AI adoption is often driven by economic incentives and competitive pressures, which can lead to underinvestment in fairness, transparency and accountability safeguards. Third, governance mechanisms remain fragmented, varying across jurisdictions, industries and organizations, making it difficult to enforce consistent ethical principles in practice.

The literature review identified recurring ethical gaps across case studies and domains. Bias and discrimination persist due to skewed training data, opaque model architectures and insufficient post-deployment monitoring. Accountability is often diluted, with unclear lines of responsibility when harm occurs. Transparency challenges, particularly the black box nature of deep learning, limit meaningful oversight, informed consent and public trust. Privacy risks are magnified by large-scale data collection and integration, especially in contexts like surveillance, predictive policing and personalized advertising. These concerns are not hypothetical; they have already manifested in real-world harms, disproportionately affecting marginalized communities.

While the challenges are significant, they are not insurmountable. Bridging the ethical gap requires a multi-layered approach that combines technical innovation, institutional reform and public engagement. From a technical standpoint, research into explainable AI, robust fairness metrics and privacy-preserving methods (such as federated learning) can reduce some risks. Organizationally, embedding ethics into the AI lifecycle, through impact assessments, interdisciplinary review boards and ethics-by-design methodologies, can ensure that potential harms are addressed before deployment. Policymakers can play a crucial role by creating flexible yet enforceable regulations that align with evolving best practices, avoiding both regulatory lag and stifling overreach.

The responsibility for ethical AI cannot rest on a single group. Developers must integrate ethical safeguards as core features, not optional add-ons. Businesses need to balance short-term market pressures with long-term reputational and societal considerations. Governments should act as both regulators and conveners, fostering dialogue among technologists, ethicists and the public. Civil society has a role in amplifying underrepresented voices and holding powerful actors accountable. Finally, the public must be empowered to engage critically with AI systems, understanding both their capabilities and their limitations.

Modern AI is at an inflection point. If the current trajectory continues unchecked, we risk deepening inequities, eroding trust in digital systems and entrenching opaque decision-making



in critical societal functions. However, with deliberate, coordinated action, AI can be steered toward outcomes that reflect shared human values. The key lies in embracing an anticipatory, precautionary mindset, building guardrails before harm occurs rather than reacting after the fact.

Ultimately, the ethical future of AI will be determined not by algorithms alone, but by the choices of those who design, deploy and govern them. The path forward demands humility, vigilance and a commitment to ensuring that technological power is exercised responsibly. By closing the ethical gaps identified in this deliverable, we can move toward AI systems that are not only powerful and efficient, but also just, transparent and aligned with the public good.

References

- Abadi, Martin, et al. "Deep Learning with Differential Privacy." *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* [Vienna Austria], 2016, pp. 308–18. *DOI.org (Crossref)*, <https://doi.org/10.1145/2976749.2978318>
- Acemoglu, Daron. "The Simple Macroeconomics of AI." *Economic Policy*, vol. 40, no. 121, Jan. 2025, pp. 13–58. *Silverchair*, <https://doi.org/10.1093/epolic/eiae042>.
- Ali, Mutahar, et al. "Understanding Users' Security and Privacy Concerns and Attitudes Towards Conversational AI Platforms." 2025, pp. 298–316. *arXiv.org*, <https://doi.org/10.1109/SP61157.2025.00241>.
- Allen, William Hand, et al. "Fuzzy Neural Network-Based Health Monitoring for HVAC System Variable-Air-Volume Unit." *IEEE Transactions on Industry Applications*, vol. 52, no. 3, May 2016, pp. 2513–24. *IEEE Xplore*, <https://doi.org/10.1109/TIA.2015.2511160>.
- Alvarez-Melis, David, and Tommi S. Jaakkola. "Towards Robust Interpretability with Self-Explaining Neural Networks." *Proceedings of the 32nd International Conference on Neural Information Processing Systems* [Red Hook, NY, USA], NIPS'18, 2018, pp. 7786–95.
- Amershi, Saleema, et al. "Guidelines for Human-AI Interaction." *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* [New York, NY, USA], CHI '19, 2019, pp. 1–13. *ACM Digital Library*, <https://doi.org/10.1145/3290605.3300233>.
- Angwin, J., et al. "Machine Bias." *ProPublica*, 2016, <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
- Arduini, Mario, et al. "Adversarial Learning for Debiasing Knowledge Graph Embeddings." *arXiv preprint arXiv:2006.16309* (2020). <https://arxiv.org/abs/2006.16309>
- Atwya, Mohamed, and George Panoutsos. "Transient Thermography for Flaw Detection in Friction Stir Welding: A Machine Learning Approach." *IEEE Transactions on Industrial Informatics*, vol. 16, no. 7, July 2020, pp. 4423–35. *IEEE Xplore*, <https://doi.org/10.1109/TII.2019.2948023>.
- Bang, Yejin, et al. "HalluLens: LLM Hallucination Benchmark." *arXiv:2504.17550*, *arXiv*, 24 Apr. 2025. *arXiv.org*, <https://doi.org/10.48550/arXiv.2504.17550>.
- Barocas, Solon, and Andrew D. Selbst. "Big Data's Disparate Impact." SSRN Scholarly Paper no. 2477899, Social Science Research Network, 2016. *papers.ssrn.com*, <https://doi.org/10.2139/ssrn.2477899>.
- Barrett, Lisa Feldman, et al. "Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements." *Psychological Science in the Public Interest*, vol. 20, no. 1, July 2019, pp. 1–68. *DOI.org (Crossref)*, <https://doi.org/10.1177/1529100619832930>.
- Bednár, Peter, et al. "Deep Learning Methods for Multilingual Classification of Business Documents." 2025 *IEEE 23rd World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, 2025, pp. 000327–30. *IEEE Xplore*, <https://doi.org/10.1109/SAMI63904.2025.10883330>.



Bender, Emily M., et al. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* [Virtual Event Canada], 2021, pp. 610–23. Crossref, <https://doi.org/10.1145/3442188.3445922>.

Bereska, Leonard, and Efstratios Gavves. “Mechanistic Interpretability for AI Safety -- A Review.” arXiv:2404.14082, arXiv, 23 Aug. 2024. *arXiv.org*, <https://doi.org/10.48550/arXiv.2404.14082>.

Bhagat, Saidarshan, et al. “Accessibility Evaluation of Major Assistive Mobile Applications Available for the Visually Impaired.” *ITU Journal on Future and Evolving Technologies*, vol. 4, no. 4, Dec. 2023, pp. 631–43. *arXiv.org*, <https://doi.org/10.52953/TNRV4696>.

Bhatt, Umang, et al. “Explainable Machine Learning in Deployment.” *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* [Barcelona Spain], 2020, pp. 648–57. DOI.org (Crossref), <https://doi.org/10.1145/3351095.3375624>.

Bi, Xiao, et al. “DeepSeek LLM: Scaling Open-Source Language Models with Longtermism.” arXiv:2401.02954, arXiv, 5 Jan. 2024. *arXiv.org*, <https://doi.org/10.48550/arXiv.2401.02954>.

Binns, Reuben. “Fairness in Machine Learning: Lessons from Political Philosophy.” *arXiv [Cs.CY]*, 2017.

Blodgett, Su Lin, et al. “Language (Technology) Is Power: A Critical Survey of ‘Bias’ in NLP.” *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by Dan Jurafsky et al., Association for Computational Linguistics, 2020, pp. 5454–76. ACLWeb, <https://doi.org/10.18653/v1/2020.acl-main.485>.

Bolukbasi, Tolga, et al. “Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings.” *NIPS: Neural Information Processing Systems*, 2016, pp. 4356–64, <https://dl.acm.org/doi/10.5555/3157382.3157584>.

Borodin, Yevgen, et al. “More than Meets the Eye: A Survey of Screen-Reader Browsing Strategies.” *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A)* [New York, NY, USA], W4A ’10, 2010, pp. 1–10. ACM Digital Library, <https://doi.org/10.1145/1805986.1806005>.

Brickley, D., and R. V. Guha. “RDF Schema 1.1.” W3C Recommendation, 2014, <https://www.w3.org/TR/rdf-schema/>.

Briggs, J., and D. Kodnani. *Generative AI Could Raise Global GDP by 7%*. 2023, <https://www.goldmansachs.com/insights/articles/generative-ai-could-raise-global-gdp-by-7-percent>.

Buchholz, Katharina. “The Extreme Cost Of Training AI Models Like ChatGPT and Gemini.” *Forbes*, 2024, <https://www.forbes.com/sites/katharinabuchholz/2024/08/23/the-extreme-cost-of-training-ai-models/>.

Budnik, Christian. “Can We Trust Artificial Intelligence?” *Philosophy & Technology*, vol. 38, no. 1, Mar. 2025, p. 10. DOI.org (Crossref), <https://doi.org/10.1007/s13347-024-00820-1>.

Buocz, Thomas, et al. “Regulatory Sandboxes in the AI Act: Reconciling Innovation and Safety?” *Law, Innovation and Technology*, vol. 15, no. 2, July 2023, pp. 357–89. Taylor and Francis+NEJM, <https://doi.org/10.1080/17579961.2023.2245678>.



Bütthe, Tim, et al. “Governing AI – Attempting to Herd Cats? Introduction to the Special Issue on the Governance of Artificial Intelligence.” *Journal of European Public Policy*, vol. 29, no. 11, Nov. 2022, pp. 1721–52. *DOI.org (Crossref)*, <https://doi.org/10.1080/13501763.2022.2126515>.

Cai, Liwei, and William Yang Wang. “KBGAN: Adversarial Learning for Knowledge Graph Embeddings.” *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, edited by Marilyn Walker et al., Association for Computational Linguistics, 2018, pp. 1470–80, <https://doi.org/10.18653/v1/N18-1133>.

Calgaro, Emma. “Silent No More: Identifying and Breaking through the Barriers That d/Deaf People Face in Responding to Hazards and Disasters.” *International Journal of Disaster Risk Reduction*, Jan. 2024. www.academia.edu, <https://doi.org/10.1016/J.IJDRR.2021.102156>.

Calvert, Simeon C., et al. *Gaps in the Control of Automated Vehicles on Roads* | *IEEE Journals & Magazine* | *IEEE Xplore*. <https://ieeexplore.ieee.org/document/8977540>. Accessed 11 June 2025.

Carbonell, Jaime R. “AI in CAI: An Artificial-Intelligence Approach to Computer-Assisted Instruction.” *IEEE Transactions on Man-Machine Systems*, vol. 11, no. 4, 2007, pp. 190–202. <https://doi.org/10.1109/TMMS.1970.299942>

Carlini, Nicholas, et al. “Extracting Training Data from Large Language Models.” 2021, pp. 2633–50. www.usenix.org, <https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting>.

Chakravarti, Suman K., and Sai Radha Mani Alla. “Descriptor Free QSAR Modeling Using Deep Learning with Long Short-Term Memory Neural Networks.” *Frontiers in Artificial Intelligence*, vol. 2, 2019, p. 17.

Chen, Jiahao, et al. “Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved.” *Proceedings of the Conference on Fairness, Accountability, and Transparency* [New York, NY, USA], FAT* ’19, 2019, pp. 339–48. *ACM Digital Library*, <https://doi.org/10.1145/3287560.3287594>.

Chen, Zhi, and Lingxiao Jiang. “Promise and Peril of Collaborative Code Generation Models: Balancing Effectiveness and Memorization.” *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering* [Sacramento CA USA], 2024, pp. 493–505. *DOI.org (Crossref)*, <https://doi.org/10.1145/3691620.3695021>.

Cheng, J., et al. “A Survey on Graph Computing: Models, Tools, and Applications.” *ACM Computing Surveys*, vol. 53, no. 5, 2020, pp. 1–38, <https://doi.org/10.1145/3402876>.

Cheng, Yong, et al. “Mu²SLAM: Multitask, Multilingual Speech and Language Models.” *Proceedings of the 40th International Conference on Machine Learning*, 2023, pp. 5504–20. proceedings.mlr.press, <https://proceedings.mlr.press/v202/cheng23e.html>.

Chin, Kuan-Chen, et al. “Early Recognition of a Caller’s Emotion in out-of-Hospital Cardiac Arrest Dispatching: An Artificial Intelligence Approach.” *Resuscitation*, vol. 167, 2021, pp. 144–50.

Chiu, Thomas K.F., et al., “Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education” *Computers and Education: Artificial Intelligence*, vol.4, 2023, <https://doi.org/10.1016/j.caeai.2022.100118>



Chowdhary, Shreya, et al. “Technical Solutions to Emotion AI’s Privacy Harms: A Systematic Literature Review.” *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* [Athens Greece], 2025, pp. 1119–44. *DOI.org (Crossref)*, <https://doi.org/10.1145/3715275.3732074>.

Chuang, Yu-Neng, et al. “Fair-RGNN: Mitigating Relational Bias on Knowledge Graphs.” *ACM Trans. Knowl. Discov. Data*, vol. 19, no. 2, Feb. 2025, pp. 1–18.

Cichońska, Anna, et al. “Crowdsourced Mapping of Unexplored Target Space of Kinase Inhibitors.” *Nature Communications*, vol. 12, no. 1, 2021, p. 3307, <https://doi.org/10.1038/s41467-021-23165-1>

Collingridge, David. *The Social Control of Technology*. Frances Pinter St. Martin’s press, 1982.

Commission, European, et al. *Ethics Guidelines for Trustworthy AI*. Publications Office, 2019, <https://doi.org/10.2759/346720>.

Conneau, Alexis, et al. “Unsupervised Cross-Lingual Representation Learning at Scale.” arXiv:1911.02116, arXiv, 8 Apr. 2020. *arXiv.org*, <https://doi.org/10.48550/arXiv.1911.02116>.

Cowls, Josh, et al. “The AI Gambit: Leveraging Artificial Intelligence to Combat Climate Change—Opportunities, Challenges, and Recommendations.” *AI & SOCIETY*, vol. 38, no. 1, Feb. 2023, pp. 283–307. *DOI.org (Crossref)*, <https://doi.org/10.1007/s00146-021-01294-x>.

Crawford, Kate, et al. *AI Now 2019 Report*. New York, 2019, <https://ainowinstitute.org/publications/ai-now-2019-report-2>.

Cui, Jian, Zirui Lan, Yisi Liu, et al. “A Compact and Interpretable Convolutional Neural Network for Cross-Subject Driver Drowsiness Detection from Single-Channel EEG.” *Methods*, vol. 202, June 2022, pp. 173–84. *arXiv.org*, <https://doi.org/10.1016/j.ymeth.2021.04.017>.

Cui, Jian, Zirui Lan, Olga Sourina, et al. “EEG-Based Cross-Subject Driver Drowsiness Recognition With an Interpretable Convolutional Neural Network.” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 10, Oct. 2023, pp. 7921–33. *IEEE Xplore*, <https://doi.org/10.1109/TNNLS.2022.3147208>.

Dai, Ling, et al. “A Deep Learning System for Predicting Time to Progression of Diabetic Retinopathy.” *Nature Medicine*, vol. 30, no. 2, 2024, pp. 584–94, <https://doi.org/10.1038/s41591-023-02702-z>

Dastin, J. “Amazon Scraps Secret AI Recruiting Tool After Finding It Was Biased Against Women.” *Reuters*, 2018, <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>.

Davies, Harry, et al. “‘The Gospel’: How Israel Uses AI to Select Bombing Targets in Gaza.” *The Guardian*, 1 Dec. 2023. World News. *The Guardian*, <https://www.theguardian.com/world/2023/dec/01/the-gospel-how-israel-uses-ai-to-select-bombing-targets>.

Deck, Luca, et al. “A Critical Survey on Fairness Benefits of Explainable AI.” *The 2024 ACM Conference on Fairness Accountability and Transparency* [Rio de Janeiro Brazil], 2024, pp. 1579–95. *DOI.org (Crossref)*, <https://doi.org/10.1145/3630106.3658990>.

- Devlin, Jacob, et al. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *arXiv*, 2018. *DOI.org (Datacite)*, <https://doi.org/10.48550/ARXIV.1810.04805>.
- Díaz-Santos, Sonia, et al. "Driver Identification and Detection of Drowsiness While Driving." *Applied Sciences*, vol. 14, no. 6, no. 6, Jan. 2024, p. 2603. *www.mdpi.com*, <https://doi.org/10.3390/app14062603>.
- Dillet, Romain. "European AI Startups Raised \$8B in 2024." *TechCrunch*, 5 Feb. 2025, <https://techcrunch.com/2025/02/04/european-ai-startups-raised-8-billion-in-2024/>.
- Dilsizian, Matthew E., and Eliot L. Siegel. "Machine Meets Biology: A Primer on Artificial Intelligence in Cardiology and Cardiac Imaging." *Current Cardiology Reports*, vol. 20, no. 12, Dec. 2018, p. 139. *DOI.org (Crossref)*, <https://doi.org/10.1007/s11886-018-1074-8>.
- Dimlioglu, Tolga, et al. "Automatic Document Classification via Transformers for Regulations Compliance Management in Large Utility Companies." *Neural Computing and Applications*, vol. 35, no. 23, Aug. 2023, pp. 17167–85. *Springer Link*, <https://doi.org/10.1007/s00521-023-08555-4>.
- Dixon, Patrick. *How AI Will Change Your Life*. Profile Books, 2024.
- Doerfler, Aaron, and Luke Stark. "Legitimizing Emotion Tracking Technologies in Driver Monitoring Systems." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 7, 2024, pp. 396–410. *Google Scholar*, <https://ojs.aaai.org/index.php/AIES/article/view/31645>.
- Dogan, Ebru, et al. "Ethical Issues Concerning Automated Vehicles and Their Implications for Transport." vol. 5, 2020, pp. 215–33. *Semantic Scholar*, <https://doi.org/10.1016/bs.atpp.2020.05.003>.
- Doshi-Velez, F., and B. Kim. "Towards a Rigorous Science of Interpretable Machine Learning." *arXiv Preprint arXiv:1702.08608*, 2017, <https://arxiv.org/abs/1702.08608>.
- Duewer, David L. *Face Recognition Vendor Test (FRVT) Part 8: Summarizing Demographic Differentials*. NIST IR 8429, National Institute of Standards and Technology, 2022, p. NIST IR 8429. *DOI.org (Crossref)*, <https://doi.org/10.6028/NIST.IR.8429>.
- Dwork, C. "Differential Privacy." *Automata, Languages and Programming. 33rd International Colloquium, ICALP 2006*, vol. 4052, 2006, pp. 35–46, https://doi.org/10.1007/11787006_35.
- Dwork, Cynthia, and Aaron Roth. "The Algorithmic Foundations of Differential Privacy." *Foundations and Trends® in Theoretical Computer Science*, vol. 9, nos. 3–4, 2014, pp. 211–407, <https://doi.org/10.1561/04000000042>.
- Ebrahimian, Serajeddin, et al. "Multi-Level Classification of Driver Drowsiness by Simultaneous Analysis of ECG and Respiration Signals Using Deep Neural Networks." *International Journal of Environmental Research and Public Health*, vol. 19, no. 17, no. 17, Jan. 2022, p. 10736. *www.mdpi.com*, <https://doi.org/10.3390/ijerph191710736>.
- Esteves, Telma, et al. "AUTOMOTIVE: A Case Study on AUTOMATIC multiMODal Drowsiness detection for Smart VEHICLES." *IEEE Access*, vol. 9, 2021, pp. 153678–700. *Semantic Scholar*, <https://doi.org/10.1109/ACCESS.2021.3128016>.

Etzioni, Oren, et al. "Open Information Extraction from the Web." *Communications of the ACM*, vol. 51, no. 12, Dec. 2008, pp. 68–74. DOI.org (Crossref), <https://doi.org/10.1145/1409360.1409378>.

European Commission: Directorate-General for Research and Innovation, Breque, M., De Nul, L. and Petridis, A., Industry 5.0 – Towards a sustainable, human-centric and resilient European industry, Publications Office of the European Union, 2021, <https://data.europa.eu/doi/10.2777/308407>.

European Parliament & Council of the European Union. *Directive 2000/78/EC establishing a general framework for equal treatment in employment and occupation* (Employment Equality Directive). *Official Journal of the European Communities*, L 303, 16-22, 2000. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32000L0078>.

European Parliament & Council of the European Union. *General Data Protection Regulation (GDPR), Regulation (EU) 2016/679*. *Official Journal of the European Union*, L 119, 1-88, 2016. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679>.

European Parliament & Council of the European Union. *Medical Devices Regulation, Regulation (EU) 2017/745*. *Official Journal of the European Union*, L 117, 1-175, 2017. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32017R0745>.

European Parliament & Council of the European Union. *Machinery Regulation, Regulation (EU) 2023/1230*. *Official Journal of the European Union*, L 165, 1-118, 2023. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32023R1230>.

European Parliament & Council of the European Union. *Artificial Intelligence Act, Regulation (EU) 2024/1689*. *Official Journal of the European Union*, L 259, 1-157, 2024. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689>.

European Union. *Charter of Fundamental Rights of the European Union* (2012/C 326/02). *Official Journal of the European Union*, C 326, 391-407, 2012. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT>.

Farzaneh, Hooman, et al. "Artificial Intelligence Evolution in Smart Buildings for Energy Efficiency." *Applied Sciences*, vol. 11, no. 2, 2021, p. 763, <https://doi.org/10.3390/app11020763>

Feldstein, Steven. "AI & Big Data Global Surveillance Index (2022 Updated)." *Mendeley Data*, vol. 3, 2022, p. 2022, <https://doi.org/10.17632/gjhf5y4xjp.4>

Fellows, M., et al. "Scaling Graph Databases for Big Data Applications." *Journal of Big Data*, vol. 8, no. 1, 2021, pp. 22–41, <https://doi.org/10.1186/s40537-021-00430-7>.

Fink, Melanie. "Human Oversight under Article 14 of the EU AI Act." SSRN, 14 Feb. 2025, doi:10.2139/ssrn.5147196.

Forzieri, Giovanni, et al. "Emerging Signals of Declining Forest Resilience under Climate Change." *Nature*, vol. 608, no. 7923, 2022, pp. 534–39, <https://doi.org/10.1038/s41586-022-04959-9>

Fossa, Fabio, et al. "Operationalizing the Ethics of Connected and Automated Vehicles: An Engineering Perspective." *International Journal of Technoethics*, vol. 13, no. 1, Feb. 2022, pp. 1–20. *Semantic Scholar*, <https://doi.org/10.4018/IJT.291553>.



Friedler, Sorelle A., et al. "A Comparative Study of Fairness-Enhancing Interventions in Machine Learning." *Proceedings of the Conference on Fairness, Accountability, and Transparency* [New York, NY, USA], FAT* '19, 2019, pp. 329–38. *ACM Digital Library*, <https://doi.org/10.1145/3287560.3287589>.

Friedman, Batya, et al. "Value Sensitive Design and Information Systems." *Early Engagement and New Technologies: Opening up the Laboratory*, edited by Neelke Doorn et al., vol. 16, Springer Netherlands, 2013, pp. 55–95. *Philosophy of Engineering and Technology. DOI.org (Crossref)*, https://doi.org/10.1007/978-94-007-7844-3_4.

Fu, Biying, et al. "A Survey on Drowsiness Detection -- Modern Applications and Methods." arXiv:2408.12990, arXiv, 23 Aug. 2024. *arXiv.org*, <https://doi.org/10.48550/arXiv.2408.12990>.

García-Alcaide, Diego Caballero, et al. "Analyzing Model Behavior for Driver Emotion Recognition and Drowsiness Detection Using Explainable Artificial Intelligence | Request PDF." *ResearchGate*, 2025. [www.researchgate.net](https://doi.org/10.5220/0013204400003941), <https://doi.org/10.5220/0013204400003941>.

Garg, Nikhil, et al. "Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes." *Proceedings of the National Academy of Sciences*, vol. 115, no. 16, Apr. 2018. *DOI.org (Crossref)*, <https://doi.org/10.1073/pnas.1720347115>.

Garlik, Steve Harris. "SPARQL 1.1 Query Language." W3C Recommendation, 2013, <https://www.w3.org/TR/sparql11-query/>.

Geng, Ruotong, et al. "Mitigating Sensitive Information Leakage in LLMs4Code through Machine Unlearning." arXiv:2502.05739, arXiv, 9 Feb. 2025. *arXiv.org*, <https://doi.org/10.48550/arXiv.2502.05739>.

Gohel, Prashant, et al. "Explainable AI: Current Status and Future Directions." arXiv:2107.07045, arXiv, 12 July 2021. *arXiv.org*, <https://doi.org/10.48550/arXiv.2107.07045>.

Goodfellow, Ian J., et al. "Explaining and Harnessing Adversarial Examples." arXiv:1412.6572, arXiv, 20 Mar. 2015. *arXiv.org*, <https://doi.org/10.48550/arXiv.1412.6572>.

Goodman, Bryce, and Seth Flaxman. "European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation.'" *AI Magazine*, vol. 38, no. 3, Sept. 2017, pp. 50–57. *arXiv.org*, <https://doi.org/10.1609/aimag.v38i3.2741>.

Grari, Vincent, et al. "Adversarial Learning for Counterfactual Fairness." *Mach. Learn.*, vol. 112, no. 3, Mar. 2023, pp. 741–63.

Grother, Patrick, et al. *Face Recognition Vendor Test (Fvrt): Part 3, Demographic Effects*. National Institute of Standards and Technology Gaithersburg, MD, 2019. *Google Scholar*, https://pages.nist.gov/frvt/reports/demographics/nistir_8280.pdf.

Grzelak, Jakub, and Martim Brandao. "The Dangers of Drowsiness Detection: Differential Performance, Downstream Impact, and Misuses." *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* [New York, NY, USA], AIES '21, 2021, pp. 525–31. *ACM Digital Library*, <https://doi.org/10.1145/3461702.3462593>.

Guo, Runfang, et al. "Development and Application of Emotion Recognition Technology — a Systematic Literature Review." *BMC Psychology*, vol. 12, no. 1, Feb. 2024, p. 95. *DOI.org (Crossref)*, <https://doi.org/10.1186/s40359-024-01581-4>.



- Guo, Yufei, et al. "Bias in Large Language Models: Origin, Evaluation, and Mitigation." Version 1, arXiv, 2024. *DOI.org (Datacite)*, <https://doi.org/10.48550/ARXIV.2411.10915>.
- Gwak, Jongseong, et al. "Early Detection of Driver Drowsiness Utilizing Machine Learning Based on Physiological Signals, Behavioral Measures, and Driving Performance." *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 1794–800. *IEEE Xplore*, <https://doi.org/10.1109/ITSC.2018.8569493>.
- Haigh, Thomas. *Artificial Intelligence Then and Now – Communications of the ACM*. 6 Jan. 2025, <https://cacm.acm.org/opinion/artificial-intelligence-then-and-now/>.
- Hasan, Md Mahmudul, et al. "Validation and Interpretation of a Multimodal Drowsiness Detection System Using Explainable Machine Learning." *Computer Methods and Programs in Biomedicine*, vol. 243, Jan. 2024, p. 107925. *ScienceDirect*, <https://doi.org/10.1016/j.cmpb.2023.107925>.
- Hasanzadeh, Fereshteh, et al. "Bias Recognition and Mitigation Strategies in Artificial Intelligence Healthcare Applications." *Npj Digital Medicine*, vol. 8, no. 1, Mar. 2025, p. 154. *www.nature.com*, <https://doi.org/10.1038/s41746-025-01503-7>.
- Haynes, Philip. *Managing Complexity in the Public Services*. 2nd ed., Routledge, 2015, <https://doi.org/10.4324/9781315816777>.
- Hill, Michael, and Frédéric Varone. *The Public Policy Process*. 8th ed., Routledge, 2021, <https://doi.org/10.4324/9781003010203>.
- Hogan, A., et al. "Knowledge Graphs: A Comprehensive Overview." *Semantic Web*, vol. 11, no. 3, 2020, pp. 215–36, <https://doi.org/10.3233/SW-200386>.
- Honnibal, Matthew, et al. *spaCy: Industrial-Strength Natural Language Processing in Python*. 2020, <https://doi.org/10.5281/zenodo.1212303>.
- Hossain, M. Shamim. "Patient State Recognition System for Healthcare Using Speech and Facial Expressions." *Journal of Medical Systems*, vol. 40, no. 12, Dec. 2016, p. 272. *DOI.org (Crossref)*, <https://doi.org/10.1007/s10916-016-0627-x>.
- Hosseini, Mohammad Mehdi, et al. "Faces of Fairness: Examining Bias in Facial Expression Recognition Datasets and Models." arXiv:2502.11049, arXiv, 16 Feb. 2025. *arXiv.org*, <https://doi.org/10.48550/arXiv.2502.11049>.
- Hovy, Dirk, and Shrimai Prabhumoye. "Five Sources of Bias in Natural Language Processing." *Language and Linguistics Compass*, vol. 15, no. 8, Aug. 2021. *Crossref*, <https://doi.org/10.1111/lnc3.12432>.
- Hssayeni, Murtadha D., et al. "Wearable Sensors for Estimation of Parkinsonian Tremor Severity during Free Body Movements." *Sensors*, vol. 19, no. 19, 2019, p. 4215, <https://www.mdpi.com/1424-8220/19/19/4215>.
- Iseri, Erkut Inan, et al. "Web Accessibility of the Cyprus Island Food Retailers' Websites." *2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, 2023, pp. 1–5. *IEEE Xplore*, <https://doi.org/10.1109/HORA58378.2023.10156791>.



- Jadczyk, Tomasz, Wojciech Wojakowski, et al. "Artificial Intelligence Can Improve Patient Management at the Time of a Pandemic: The Role of Voice Technology." *Journal of Medical Internet Research*, vol. 23, no. 5, 2021, p. e22959, <https://doi.org/10.2196/22959>
- Jadczyk, Tomasz, Oskar Kiwic, et al. "Feasibility of a Voice-Enabled Automated Platform for Medical Data Collection: CardioCube." *International Journal of Medical Informatics*, vol. 129, 2019, pp. 388–93, <https://doi.org/10.1016/j.ijmedinf.2019.07.001>
- Jamthe, Sudha, et al. "Inclusive Ethical AI in Human–Computer Interaction in Autonomous Vehicles." *Journal of AI, Robotics & Workplace Automation*, vol. 1, no. 3, Mar. 2022, p. 294. *Semantic Scholar*, <https://doi.org/10.69554/AMVA3377>.
- Ji, Ziwei, et al. "Survey of Hallucination in Natural Language Generation." *ACM Computing Surveys*, vol. 55, no. 12, Dec. 2023, pp. 1–38. *DOI.org (Crossref)*, <https://doi.org/10.1145/3571730>.
- Jiang, Yuchen, et al. "Quo Vadis Artificial Intelligence?" *Discover Artificial Intelligence*, vol. 2, no. 1, Mar. 2022, p. 4. *DOI.org (Crossref)*, <https://doi.org/10.1007/s44163-022-00022-8>.
- Jin, Haolin, et al. "Towards Advancing Code Generation with Large Language Models: A Research Roadmap." *arXiv:2501.11354*, arXiv, 20 Jan. 2025. *arXiv.org*, <https://doi.org/10.48550/arXiv.2501.11354>.
- Junior, Luiz Carlos da Silva Garcia, et al. "Computer Vision in Industry: An Applied Study on Autonomous Visual: Inspection of Industrial Processes." *Cuadernos de Educación y Desarrollo*, vol. 17, no. 7, 2025, pp. e8817–e8817.
- Kaffee, Lucie-Aimée, et al. "Thorny Roses: Investigating the Dual Use Dilemma in Natural Language Processing." *Findings of the Association for Computational Linguistics: EMNLP 2023* [Singapore], 2023, pp. 13977–98. *Crossref*, <https://doi.org/10.18653/v1/2023.findings-emnlp.932>.
- Kalluri, Pratyusha. "Don't Ask If Artificial Intelligence Is Good or Fair, Ask How It Shifts Power." *Nature*, vol. 583, no. 7815, July 2020, pp. 169–169. *www.nature.com*, <https://doi.org/10.1038/d41586-020-02003-2>.
- Kalusivalingam, Aravind Kumar, et al. "Enhancing Corporate Governance and Compliance through AI: Implementing Natural Language Processing and Machine Learning Algorithms." *International Journal of AI and ML*, vol. 3, no. 9, no. 9, Feb. 2022. <https://cognitivecomputingjournal.com/index.php/IJAIML-V1/article/view/73>.
- Kamalov, Firuz, et al. "New Era of Artificial Intelligence in Education: Towards a Sustainable Multifaceted Revolution." *Sustainability*, vol. 15, no. 16, 2023, p. 12451, <https://doi.org/10.48550/arXiv.2305.18303>
- Karatzia, Loucia, et al. "Artificial Intelligence in Cardiology: Hope for the Future and Power for the Present." *Frontiers in Cardiovascular Medicine*, vol. 9, 2022, p. 945726, <https://doi.org/10.3389/fcvm.2022.945726>
- Kasban, Hany, et al. "A Comparative Study of Medical Imaging Techniques." *International Journal of Information Science and Intelligent System*, vol. 4, no. 2, 2015, pp. 37–58.

Khalil, Hady A., et al. "Low-Cost Driver Monitoring System Using Deep Learning." *IEEE Access*, vol. 13, 2025, pp. 14151–64. *IEEE Xplore*, <https://doi.org/10.1109/ACCESS.2025.3530296>.

Kielland, Anders, et al. *Machine Learning Advancements for Vehicle Safety Systems*. no. 2069/2024, Institute of Transport Economics, 2024.

Kilian, Robert, et al. "European AI Standards - Technical Standardization and Implementation Challenges under the EU AI Act." SSRN Scholarly Paper no. 5155591, Social Science Research Network, 26 Feb. 2025. *papers.ssrn.com*, <https://doi.org/10.2139/ssrn.5155591>.

Kim, Dohyoung, et al. "Overview of Fair Federated Learning for Fairness and Privacy Preservation." *Expert Systems with Applications*, vol. 293, Dec. 2025, p. 128568. *ScienceDirect*, <https://doi.org/10.1016/j.eswa.2025.128568>.

Kirchenbauer, John, et al. "A Watermark for Large Language Models." arXiv, 2024, <https://arxiv.org/abs/2301.10226>

Kiritchenko, Svetlana, and Saif Mohammad. "Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems." *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, edited by Malvina Nissim et al., Association for Computational Linguistics, 2018, pp. 43–53. *ACLWeb*, <https://doi.org/10.18653/v1/S18-2005>.

Kissinger, Henry A., et al. *The Age of AI: And Our Human Future*. Little, Brown and Company, 2021.

Kleinberg, Jon, et al. "Inherent Trade-Offs in the Fair Determination of Risk Scores." arXiv:1609.05807, arXiv, 17 Nov. 2016. *arXiv.org*, <https://doi.org/10.48550/arXiv.1609.05807>.

Klijn, E. H, and J. F. M Koppenjan. "Public Management and Policy Networks: Foundations of a Network Approach to Governance." *Public Management: An International Journal of Research and Theory*, vol. 2, no. 2, 2000, pp. 135–58, <https://doi.org/10.1080/14719030000000007>.

Kotter, Elmar, et al. "Guiding AI in Radiology: ESR's Recommendations for Effective Implementation of the European AI Act." *Insights into Imaging*, vol. 16, no. 1, Feb. 2025, p. 33. *Springer Link*, <https://doi.org/10.1186/s13244-025-01905-x>.

Koulaouzidis, George, et al. "Artificial Intelligence in Cardiology—A Narrative Review of Current Status." *Journal of Clinical Medicine*, vol. 11, no. 13, Jan. 2022, p. 3910. *www.mdpi.com*, <https://doi.org/10.3390/jcm11133910>.

Kozik, Rafał, et al. "Towards Explainable Fake News Detection and Automated Content Credibility Assessment: Polish Internet and Digital Media Use-Case." *Neurocomputing*, vol. 608, Dec. 2024, p. 128450. *DOI.org (Crossref)*, <https://doi.org/10.1016/j.neucom.2024.128450>.

Kraft, Angelie, and Ricardo Usbeck. "The Lifecycle of 'Facts': A Survey of Social Bias in Knowledge Graphs." *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, edited by Yulan He et al., Association for Computational Linguistics, 2022, pp. 639–52, <https://doi.org/10.18653/v1/2022.aacl-main.49>.



Krarup, Troels, and Maja Horst. "European Artificial Intelligence Policy as Digital Single Market Making." *Big Data & Society*, vol. 10, no. 1, Jan. 2023, p. 20539517231153811. *SAGE Journals*, <https://doi.org/10.1177/20539517231153811>.

Król, Karol, and Dariusz Zdonek. "Local Government Website Accessibility—Evidence from Poland." *Administrative Sciences*, vol. 10, no. 2, no. 2, June 2020, p. 22. *www.mdpi.com*, <https://doi.org/10.3390/admsci10020022>.

Krontiris, Ioannis, et al. "Autonomous Vehicles: Data Protection and Ethical Considerations." *Proceedings of the 4th ACM Computer Science in Cars Symposium* [New York, NY, USA], CSCS '20, 2020, pp. 1–10. *ACM Digital Library*, <https://doi.org/10.1145/3385958.3430481>.

Kumar, Vivekanandan, and David Boulanger. "Explainable Automated Essay Scoring: Deep Learning Really Has Pedagogical Value." *Frontiers in Education*, vol. 5, Oct. 2020. *Frontiers*, <https://doi.org/10.3389/feduc.2020.572367>.

Latour, Bruno. *Science in Action: How to Follow Scientists and Engineers through Society*. Harvard University Press, 1987.

Li, Kangji, et al. "Forecasting Building Energy Consumption Using Neural Networks and Hybrid Neuro-Fuzzy System: A Comparative Study." *Energy and Buildings*, vol. 43, no. 10, Oct. 2011, pp. 2893–99. *ScienceDirect*, <https://doi.org/10.1016/j.enbuild.2011.07.010>.

Li, Qiong, et al. "Applying Support Vector Machine to Predict Hourly Cooling Load in the Building." *Applied Energy*, vol. 86, no. 10, Oct. 2009, pp. 2249–56. *ScienceDirect*, <https://doi.org/10.1016/j.apenergy.2008.11.035>.

Li, Wenbo, et al. "Review and Perspectives on Human Emotion for Connected Automated Vehicles." *Automotive Innovation*, vol. 7, no. 1, Feb. 2024, pp. 4–44. *DOI.org (Crossref)*, <https://doi.org/10.1007/s42154-023-00270-z>.

Li, Xinyue, et al. "Bias behind the Wheel: Fairness Testing of Autonomous Driving Systems." *ACM Trans. Softw. Eng. Methodol.*, vol. 34, no. 3, Feb. 2025, p. 82:1-82:24. *ACM Digital Library*, <https://doi.org/10.1145/3702989>.

Lin, Chien-Chang, et al. "Artificial Intelligence in Intelligent Tutoring Systems toward Sustainable Education: A Systematic Review." *Smart Learning Environments*, vol. 10, no. 1, Aug. 2023, p. 41. *Springer Link*, <https://doi.org/10.1186/s40561-023-00260-y>.

Lin, Xuan, et al. "KGNN: Knowledge Graph Neural Network for Drug-Drug Interaction Prediction." *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence* [Yokohama, Japan], 2020, pp. 2739–45. *DOI.org (Crossref)*, <https://doi.org/10.24963/ijcai.2020/380>.

Liu, Haoyang. "The Evolution of Machine Learning in Natural Language Processing: From Traditional Methods to Deep Learning." *Applied and Computational Engineering*, vol. 157, July 2025, pp. 124–31. *direct.ewa.pub*, <https://doi.org/10.54254/2755-2721/2025.PO24675>.

Liu, Siyi, et al. "Large Language Model Agent for Hyper-Parameter Optimization." *arXiv:2402.01881*, arXiv, 26 Feb. 2025. *arXiv.org*, <https://doi.org/10.48550/arXiv.2402.01881>.

Lu, Juan, et al. "Performance of Multilabel Machine Learning Models and Risk Stratification Schemas for Predicting Stroke and Bleeding Risk in Patients with Non-Valvular Atrial



Fibrillation.” *Computers in Biology and Medicine*, vol. 150, Nov. 2022, p. 106126. *ScienceDirect*, <https://doi.org/10.1016/j.combiomed.2022.106126>.

Luo, Chu Fei, et al. *BiasKG: Adversarial Knowledge Graphs to Induce Bias in Large Language Models*. 2025.

Ma, Yongfeng. “A Study of Ethical Issues in Natural Language Processing with Artificial Intelligence.” *Journal of Computer Science and Technology Studies*, vol. 5, no. 1, no. 1, Mar. 2023, pp. 52–56. *al-kindipublisher.com*, <https://doi.org/10.32996/jcsts.2023.5.1.7>.

Madsen, Andreas, et al. “Post-Hoc Interpretability for Neural NLP: A Survey.” *ACM Computing Surveys*, vol. 55, no. 8, Aug. 2023, pp. 1–42. *DOI.org (Crossref)*, <https://doi.org/10.1145/3546577>.

Maggetti, Martino. “Introduction to Regulation and Governance.” *Books*, 2025. *ideas.repec.org*, <https://ideas.repec.org/b/elg/eebook/22026.html>.

Maiti, Soumi, et al. “VoxLM: Unified Decoder-Only Models for Consolidating Speech Recognition, Synthesis and Speech, Text Continuation Tasks.” *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 13326–30. *IEEE Xplore*, <https://doi.org/10.1109/ICASSP48485.2024.10447112>.

Maleki Varnosfaderani, Shiva, and Mohamad Forouzanfar. “The Role of AI in Hospitals and Clinics: Transforming Healthcare in the 21st Century.” *Bioengineering*, vol. 11, no. 4, Apr. 2024, p. 337. *www.mdpi.com*, <https://doi.org/10.3390/bioengineering11040337>.

Manning, C. D., et al. “The Stanford CoreNLP Natural Language Processing Toolkit.” *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, 2014, pp. 55–60, <https://aclanthology.org/P14-5010/>.

Mathew, Daniel Enemona, et al. “Recent Emerging Techniques in Explainable Artificial Intelligence to Enhance the Interpretable and Understanding of AI Models for Human.” *Neural Processing Letters*, vol. 57, no. 1, Feb. 2025, p. 16. *Springer Link*, <https://doi.org/10.1007/s11063-025-11732-2>.

Mattu, Jeff Larson, Julia Angwin, Lauren Kirchner, Surya. “How We Analyzed the COMPAS Recidivism Algorithm.” *ProPublica*, <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>. Accessed 30 July 2025.

Meadows, Donella H. *Thinking in Systems: A Primer*. Earthscan, 2009.

Mehrabi, Ninareh, et al. “A Survey on Bias and Fairness in Machine Learning.” *ACM Comput. Surv.*, vol. 54, no. 6, July 2021, p. 115:1-115:35. *ACM Digital Library*, <https://doi.org/10.1145/3457607>.

Memarian, Bahar, and Tenzin Doleck. “Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and Higher Education: A Systematic Review.” *Computers and Education: Artificial Intelligence*, vol. 5, no. 100152, 2023, p. 100152.

Mienye, E., et al. “Deep Learning in Finance: A Survey of Applications and Techniques.” *AI*, vol. 5, no. 4, 2024, article 4. MDPI, <https://doi.org/10.3390/ai5040101>.

Mienye, Ibomoiye Domor, and Theo G. Swart. “A Comprehensive Review of Deep Learning: Architectures, Recent Advances, and Applications.” *Information*, vol. 15, no. 12, Dec. 2024, p. 755. *www.mdpi.com*, <https://doi.org/10.3390/info15120755>.



Mikołajewska, Emilia, et al. "Generative AI in AI-Based Digital Twins for Fault Diagnosis for Predictive Maintenance in Industry 4.0/5.0." *Applied Sciences*, vol. 15, no. 6, 2025, p. 3166, <https://doi.org/10.3390/app15063166>.

Miller, Tim. "Explanation in Artificial Intelligence: Insights from the Social Sciences." *Artificial Intelligence*, vol. 267, Feb. 2019, pp. 1–38. *ScienceDirect*, <https://doi.org/10.1016/j.artint.2018.07.007>.

Mirindi, D., Khang, A., Mirindi, F. (2025). Artificial Intelligence (AI) and Automation for Driving Green Transportation Systems: A Comprehensive Review. In: Khang, A. (eds) Driving Green Transportation System Through Artificial Intelligence and Automation. Lecture Notes in Intelligent Transportation and Infrastructure. Springer, Cham. https://doi.org/10.1007/978-3-031-72617-0_1

Mittelstadt, Brent, et al. "Explaining Explanations in AI." *Proceedings of the Conference on Fairness, Accountability, and Transparency* [New York, NY, USA], FAT* '19, 2019, pp. 279–88. *ACM Digital Library*, <https://doi.org/10.1145/3287560.3287574>.

Moulds, S., et al. "Skillful Decadal Flood Prediction." *Geophysical Research Letters*, vol. 50, no. 3, 2023, p. e2022GL100650. *Wiley Online Library*, <https://doi.org/10.1029/2022GL100650>.

Mumuni, Alhassan, and Fuseini Mumuni. "Large Language Models for Artificial General Intelligence (AGI): A Survey of Foundational Principles and Approaches." arXiv:2501.03151, arXiv, 6 Jan. 2025. *arXiv.org*, <https://doi.org/10.48550/arXiv.2501.03151>.

Munsif, Muhammad, et al. "Monitoring Neurological Disorder Patients via Deep Learning Based Facial Expressions Analysis." *Artificial Intelligence Applications and Innovations. AIAI 2022 IFIP WG 12.5 International Workshops*, edited by Ilias Maglogiannis et al., Springer International Publishing, 2022, pp. 412–23. *Springer Link*, https://doi.org/10.1007/978-3-031-08341-9_33.

Nahavandi, Darius, et al. "Application of Artificial Intelligence in Wearable Devices: Opportunities and Challenges." *Computer Methods and Programs in Biomedicine*, vol. 213, Jan. 2022, p. 106541. *ScienceDirect*, <https://doi.org/10.1016/j.cmpb.2021.106541>.

Nannini, Luca, et al. "Mapping the Landscape of Ethical Considerations in Explainable AI Research." *Ethics and Information Technology*, vol. 26, no. 3, June 2024, p. 44. *Springer Link*, <https://doi.org/10.1007/s10676-024-09773-7>.

Nasr, Milad, et al. "Scalable Extraction of Training Data from (Production) Language Models." arXiv, 2023. *DOI.org (Datacite)*, <https://doi.org/10.48550/ARXIV.2311.17035>.

Nastoska, Aleksandra, et al. "Evaluating Trustworthiness in AI: Risks, Metrics, and Applications Across Industries." *Electronics*, vol. 14, no. 13, Jan. 2025, p. 2717. *www.mdpi.com*, <https://doi.org/10.3390/electronics14132717>.

Nevo, Sella, et al. "Flood Forecasting with Machine Learning Models in an Operational Framework." *Hydrology and Earth System Sciences*, vol. 26, no. 15, Aug. 2022, pp. 4013–32. *Copernicus Online Journals*, <https://doi.org/10.5194/hess-26-4013-2022>.

Ngxande, Mkhusele, et al. "Detecting Inter-Sectional Accuracy Differences in Driver Drowsiness Detection Algorithms." *2020 International SAUPEC/RobMech/PRASA Conference*, 2020, pp. 1–6. *IEEE Xplore*, <https://doi.org/10.1109/SAUPEC/RobMech/PRASA48453.2020.9041105>.



- Nikitas, Alexandros, et al. "Artificial Intelligence, Transport and the Smart City: Definitions and Dimensions of a New Mobility Era." *Sustainability*, vol. 12, no. 7, Jan. 2020, p. 2789. www.mdpi.com, <https://doi.org/10.3390/su12072789>.
- Niu, Liang, et al. "CodexLeaks: Privacy Leaks from Code Generation Language Models in GitHub Copilot." 2023, pp. 2133–50. www.usenix.org, <https://www.usenix.org/conference/usenixsecurity23/presentation/niu>.
- Notovich, Aviv, et al. "Explainable Artificial Intelligence (XAI): Motivation, Terminology, and Taxonomy." *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook*, edited by Lior Rokach et al., Springer International Publishing, 2023, pp. 971–85. *Springer Link*, https://doi.org/10.1007/978-3-031-24628-9_41.
- Novelli, Claudio, et al. "A Robust Governance for the AI Act: AI Office, AI Board, Scientific Panel, and National Authorities." *European Journal of Risk Regulation*, Sept. 2024, pp. 1–25. *DOI.org (Crossref)*, <https://doi.org/10.1017/err.2024.57>.
- Nvidia, et al. "Nemotron-4 340B Technical Report." arXiv, 2024. *DOI.org (Datacite)*, <https://doi.org/10.48550/ARXIV.2406.11704>.
- O'Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, 2016.
- Ong, Desmond C. "An Ethical Framework for Guiding the Development of Affectively-Aware Artificial Intelligence." *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2021, pp. 1–8. *IEEE Xplore*, <https://doi.org/10.1109/ACII52823.2021.9597441>.
- OpenAI, et al. "GPT-4 Technical Report." arXiv:2303.08774, arXiv, 4 Mar. 2024. arXiv.org, <https://doi.org/10.48550/arXiv.2303.08774>.
- Oueida, Soraia, et al. *A Fair and Ethical Healthcare Artificial Intelligence System for Monitoring Driver Behavior and Preventing Road Accidents* | *SpringerLink*. https://link.springer.com/chapter/10.1007/978-3-030-89880-9_33. Accessed 10 June 2025.
- Pahde, Frederik, et al. "Reveal to Revise: An Explainable AI Life Cycle for Iterative Bias Correction of Deep Models." *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, edited by Hayit Greenspan et al., Springer Nature Switzerland, 2023, pp. 596–606. *Springer Link*, https://doi.org/10.1007/978-3-031-43895-0_56.
- Pan, Yuchen, et al. "The Impacts of Connected Autonomous Vehicles on Mixed Traffic Flow: A Comprehensive Review." *Physica A: Statistical Mechanics and Its Applications*, vol. 635, Feb. 2024, p. 129454. *ScienceDirect*, <https://doi.org/10.1016/j.physa.2023.129454>.
- Papadimitriou, Eleonora, et al. "Towards Common Ethical and Safe 'Behaviour' Standards for Automated Vehicles." *Accident Analysis & Prevention*, vol. 174, Sept. 2022, p. 106724. *Semantic Scholar*, <https://doi.org/10.1016/j.aap.2022.106724>.
- Passalacqua, Mario, et al. "Human-Centred AI in Industry 5.0: A Systematic Review." *International Journal of Production Research*, vol. 63, no. 7, 2025, pp. 2638–69, <https://doi.org/10.1080/00207543.2024.2406021>



Patel, Bhavik N., et al. "Human–Machine Partnership with Artificial Intelligence for Chest Radiograph Diagnosis." *Npj Digital Medicine*, vol. 2, no. 1, Nov. 2019, p. 111. www.nature.com, <https://doi.org/10.1038/s41746-019-0189-7>.

Peng, Hao, et al. "Differentially Private Federated Knowledge Graphs Embedding." *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021, pp. 1416–25.

Perez, Fábio, and Ian Ribeiro. "Ignore Previous Prompt: Attack Techniques For Language Models." arXiv:2211.09527, arXiv, 17 Nov. 2022. arXiv.org, <https://doi.org/10.48550/arXiv.2211.09527>.

Pérez, J., et al. "Graph Databases for Knowledge Graphs: A Survey of Tools and Techniques." *Journal of Knowledge Management*, vol. 22, no. 5, 2018, pp. 1103–22, <https://doi.org/10.1108/JKM-06-2017-0244>.

Perrotte, Gaëtan, et al. "Monitoring Driver Drowsiness in Partially Automated Vehicles: Added Value from Combining Postural and Physiological Indicators." *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 100, Jan. 2024, pp. 458–74. *ScienceDirect*, <https://doi.org/10.1016/j.trf.2023.12.010>.

Pham, Bao-Chau, and Sarah R. Davies. "What Problems Is the AI Act Solving? Technological Solutionism, Fundamental Rights, and Trustworthiness in European AI Policy." *Critical Policy Studies*, vol. 19, no. 2, Apr. 2025, pp. 318–36. *DOI.org (Crossref)*, <https://doi.org/10.1080/19460171.2024.2373786>.

Pikuliak, Matúš, et al. "Women Are Beautiful, Men Are Leaders: Gender Stereotypes in Machine Translation and Language Modeling." arXiv:2311.18711, arXiv, 30 Sept. 2024. arXiv.org, <https://doi.org/10.48550/arXiv.2311.18711>.

Plathottam, Siby Jose, et al. "A Review of Artificial Intelligence Applications in Manufacturing Operations." *Journal of Advanced Manufacturing and Processing*, vol. 5, no. 3, 2023, p. e10159. *Wiley Online Library*, <https://doi.org/10.1002/amp2.10159>.

Pribeanu, Costin, et al. *Web Accessibility in Romania: The Conformance of Municipal Web Sites to Web Content Accessibility Guidelines*. vol. 16, no. 1.

Purnamasari, Prima Dewi, et al. "EEG Based Patient Emotion Monitoring Using Relative Wavelet Energy Feature and Back Propagation Neural Network." *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015, pp. 2820–23. *IEEE Xplore*, <https://doi.org/10.1109/EMBC.2015.7318978>.

Purohit, Jahanvi, et al. "An Artificial Intelligence Based Prototype of Driver Drowsiness Detection for Intelligent Vehicles." *2023 IEEE World AI IoT Congress (AllIoT)* [Seattle, WA, USA], June 2023, pp. 0633–40. *2023 IEEE World AI IoT Congress (AllIoT)*, *Semantic Scholar*, <https://doi.org/10.1109/AllIoT58121.2023.10174507>.

Qamar, Md. Tauseef, et al. "The Language of Nuance: Exploring the Limits of Large Language Models in Handling Ambiguity." *Big Data Analytics in Astronomy, Science, and Engineering*, edited by Shelly Sachdeva et al., Springer Nature Switzerland, 2025, pp. 180–92. *Springer Link*, https://doi.org/10.1007/978-3-031-86193-2_12.



Radford, Alec, et al. "Learning Transferable Visual Models From Natural Language Supervision." arXiv:2103.00020, arXiv, 26 Feb. 2021. *arXiv.org*, <https://doi.org/10.48550/arXiv.2103.00020>.

Rawashdeh, Awni. "The Consequences of Artificial Intelligence: An Investigation into the Impact of AI on Job Displacement in Accounting | Journal of Science and Technology Policy Management | Emerald Publishing." *Journal of Science and Technology Policy Management*, vol. 16, no. 3, Mar. 2025, pp 506–535. <https://doi.org/10.1108/JSTPM-02-2023-0030>

Rhodes, R. A. "Understanding governance: Policy networks, governance, reflexivity and accountability", 1997. Maidenhead: Open University.

Rhodes, R. A. W. "The Governance Narrative: Key Findings and Lessons from the Erc's Whitehall Programme." *Public Administration*, vol. 78, Jan. 2000, pp. 345–63. *ResearchGate*, <https://doi.org/10.1111/1467-9299.00209>.

Rhodes, R.A. "The Governance Narrative: Key Findings and Lessons from the ESRC's Whitehall Programme." Public Management and Policy Association., 2000, <https://onlinelibrary.wiley.com/doi/epdf/10.1111/1467-9299.00209>.

Ribeiro, Marco Tulio, et al. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* [New York, NY, USA], KDD '16, 2016, pp. 1135–44. *ACM Digital Library*, <https://doi.org/10.1145/2939672.2939778>.

Ribera, Mireia, et al. "Impact of Accessibility Barriers on the Mood of Users with Motor and Dexterity Impairments." *Journal of Accessibility and Design for All*, vol. 5, no. 1, no. 1, May 2015, pp. 1–26. *www.jaccs.org*, <https://doi.org/10.17411/jaccs.v5i1.93>.

Rickman, Sam. "Evaluating Gender Bias in Large Language Models in Long-Term Care." *BMC Medical Informatics and Decision Making*, vol. 25, no. 1, Aug. 2025, p. 274. *BioMed Central*, <https://doi.org/10.1186/s12911-025-03118-0>.

Righi, Riccardo, et al. "AI Watch Index 2021." *JRC Publications Repository*, 2022, <https://doi.org/10.2760/921564>.

Robinson, Ian, et al. *Graph Databases: New Opportunities for Connected Data*. O'Reilly Media, 2015.

Rohlinger, Tihomir, et al. "Deep Learning-Based EEG Detection of Mental Alertness States from Drivers under Ethical Aspects." *2021 The 5th International Conference on Advances in Artificial Intelligence (ICAAI)* [Virtual Event United Kingdom], Nov. 2021, pp. 54–64. *ICAAI 2021: 2021 the 5th International Conference on Advances in Artificial Intelligence*, *Semantic Scholar*, <https://doi.org/10.1145/3505711.3505719>.

Saadatnejad, Saeed, et al. "LSTM-Based ECG Classification for Continuous Monitoring on Personal Wearable Devices." *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 2, Feb. 2020, pp. 515–23. *IEEE Xplore*, <https://doi.org/10.1109/JBHI.2019.2911367>.

Sahayadhas, Arun, et al. "Detecting Driver Drowsiness Based on Sensors: A Review." *Sensors*, vol. 12, no. 12, no. 12, Dec. 2012, pp. 16937–53. *www.mdpi.com*, <https://doi.org/10.3390/s121216937>.



Sainath, Tara N., et al. "JOIST: A Joint Speech and Text Streaming Model for ASR." *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 52–59. *IEEE Xplore*, <https://doi.org/10.1109/SLT54892.2023.10022774>.

Saleem, Adil Ali, et al. "A Systematic Review of Physiological Signals Based Driver Drowsiness Detection Systems." *Cognitive Neurodynamics*, vol. 17, no. 5, Oct. 2023, pp. 1229–59. *PubMed*, <https://doi.org/10.1007/s11571-022-09898-9>.

Samy Abd, El-Nabi, et al. "Machine Learning and Deep Learning Techniques for Driver Fatigue and Drowsiness Detection: A Review." *ResearchGate*, Dec. 2024. www.researchgate.net, <https://doi.org/10.1007/s11042-023-15054-0>.

Santoni de Sio, Filippo. "The European Commission Report on Ethics of Connected and Automated Vehicles and the Future of Ethics of Transportation." *Ethics and Information Technology*, vol. 23, no. 4, Dec. 2021, pp. 713–26. *Springer Link*, <https://doi.org/10.1007/s10676-021-09609-8>.

Schlichtkrull, Michael, et al. "Modeling Relational Data with Graph Convolutional Networks." *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings* [Berlin, Heidelberg], 2018, pp. 593–607, https://doi.org/10.1007/978-3-319-93417-4_38.

Schneider, Johannes, et al. "Foundation Models." *Business & Information Systems Engineering*, vol. 66, no. 2, Apr. 2024, pp. 221–31. *Springer Link*, <https://doi.org/10.1007/s12599-024-00851-0>.

Schröder, Markus, et al. "A Human-in-the-Loop Approach for Personal Knowledge Graph Construction from File Names." *KGCW@ ESWC*, 2022.

Shultz, Thomas R., et al. "Text Understanding in GPT-4 versus Humans." *Royal Society Open Science*, vol. 12, no. 2, Feb. 2025, p. 241313. royalsocietypublishing.org (Atypon), <https://doi.org/10.1098/rsos.241313>.

Singh, Ninni, et al. "SeisTutor: A Custom-Tailored Intelligent Tutoring System and Sustainable Education." *Sustainability*, vol. 14, no. 7, Jan. 2022, p. 4167. www.mdpi.com, <https://doi.org/10.3390/su14074167>.

Slater, Louise J., et al. "Hybrid Forecasting: Blending Climate Predictions with AI Models." *Hydrology and Earth System Sciences*, vol. 27, no. 9, May 2023, pp. 1865–89. *Copernicus Online Journals*, <https://doi.org/10.5194/hess-27-1865-2023>.

Smuha, Nathalie A., and Karen Yeung. "The European Union's AI Act: Beyond Motherhood and Apple Pie?" *The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence*, edited by Nathalie A. Smuha, Cambridge University Press, 2025, pp. 228–58. *Cambridge Law Handbooks*. *Cambridge University Press*, <https://doi.org/10.1017/9781009367783.015>.

Solaymani, Saeed. "CO2 Emissions Patterns in 7 Top Carbon Emitter Economies: The Case of Transport Sector." *Energy*, vol. 168, Feb. 2019, pp. 989–1001. *ScienceDirect*, <https://doi.org/10.1016/j.energy.2018.11.145>.

Solove, Daniel. "A Taxonomy of Privacy." *University of Pennsylvania Law Review*, vol. 154, no. 3, Jan. 2006, p. 477.

Sousa e Silva, N.. "The Artificial Intelligence Act: Critical Overview." *JIPITEC – Journal of Intellectual Property, Information Technology and E-Commerce Law*, vol. 16, no. 1, no. 1, Mar. 2025. www.jipitec.eu, <https://www.jipitec.eu/jipitec/article/view/418>.

Sreeharsha, Apparaju, et al. "Towards Data-Driven Hydration Monitoring: Insights from Wearable Sensors and Advanced Machine Learning Techniques." *Electronics*, vol. 13, no. 24, Jan. 2024, p. 4960. www.mdpi.com, <https://doi.org/10.3390/electronics13244960>.

Stewart, Leo. "Exploring How Ai And Machine Learning Can Be Applied In Compliance To Detect Anomalies And Predict Compliance Risks." SSRN Scholarly Paper no. 5249005, Social Science Research Network, 21 Feb. 2025. papers.ssrn.com, <https://doi.org/10.2139/ssrn.5249005>.

Strubell, Emma, et al. "Energy and Policy Considerations for Deep Learning in NLP." arXiv:1906.02243, arXiv, 5 June 2019. arXiv.org, <https://doi.org/10.48550/arXiv.1906.02243>.

Suresh, Harini, and John V. Gutttag. "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle." *Equity and Access in Algorithms, Mechanisms, and Optimization*, 2021, pp. 1–9. arXiv.org, <https://doi.org/10.1145/3465416.3483305>.

Sütfeld, Leon, et al. "Towards a Framework for Ethical Decision Making in Automated Vehicles | Request PDF." *ResearchGate*, <https://doi.org/10.31234/osf.io/4duca>. Accessed 11 June 2025.

Swyngedouw, Erik. "Capitalism Is over, but the New Is Worse: Reflections on Rent, Services, and Capitalist Feudalism." *The Value of Place*, Edward Elgar Publishing, 2025, pp. 23–39. www.elgaronline.com, <https://www.elgaronline.com/edcollchap/book/9781035347926/chapter2.xml>.

Tallberg, Jonas, et al. "AI Regulation in the European Union: Examining Non-State Actor Preferences." *Business and Politics*, vol. 26, no. 2, June 2024, pp. 218–39. *Cambridge University Press*, <https://doi.org/10.1017/bap.2023.36>.

Tang, Xiaobin, et al. "BERT-INT: A BERT-Based Interaction Model for Knowledge Graph Alignment." *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence* [Yokohama, Yokohama, Japan], IJCAI'20, 2021.

Thalpage, Nipuna Sankalpa, et al. "Ethical Challenges in Explainable AI: A Review on Cultural and Social Bias - Institute of Cited Scientists." *Journal of Digital Art & Humanities*, vol. 6, no. 1, 2025, p. 30–39. ics.events, https://doi.org/10.33847/2712-8148.6.1_3.

Thornton, Sarah M., et al. "Incorporating Ethical Considerations Into Automated Vehicle Control." *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 6, June 2017, pp. 1429–39. *Semantic Scholar*, <https://doi.org/10.1109/TITS.2016.2609339>.

Tiddi, Ilaria, et al. *Knowledge Graphs for eXplainable Artificial Intelligence: Foundations, Applications, and Challenges*. IOS Press, 2020, <https://www.iospress.com/catalog/books/knowledge-graphs-for-explainable-artificial-intelligence-foundations-applications-and>. Studies on the Semantic Web.



- Toosi, Amirhosein, et al. "A Brief History of AI: How to Prevent Another Winter (A Critical Review)." *PET Clinics*, vol. 16, no. 4, Oct. 2021, pp. 449–69. www.pet.theclinics.com, <https://doi.org/10.1016/j.cpet.2021.07.001>.
- Tornede, Alexander, et al. "AutoML in the Age of Large Language Models: Current Challenges, Future Opportunities and Risks." arXiv, 2023. *DOI.org (Datacite)*, <https://doi.org/10.48550/ARXIV.2306.08107>.
- Touvron, Hugo, et al. "Llama 2: Open Foundation and Fine-Tuned Chat Models." arXiv, 2023. *DOI.org (Datacite)*, <https://doi.org/10.48550/ARXIV.2307.09288>.
- Trirat, Patara, et al. "AutoML-Agent: A Multi-Agent LLM Framework for Full-Pipeline AutoML." arXiv:2410.02958, arXiv, 6 June 2025. *arXiv.org*, <https://doi.org/10.48550/arXiv.2410.02958>.
- Tsai, Yun-Da, et al. "AutoML-GPT: Large Language Model for AutoML." arXiv:2309.01125, arXiv, 3 Sept. 2023. *arXiv.org*, <https://doi.org/10.48550/arXiv.2309.01125>.
- Tsanas, A., et al. "Daily Longitudinal Self-Monitoring of Mood Variability in Bipolar Disorder and Borderline Personality Disorder." *Journal of Affective Disorders*, vol. 205, Nov. 2016, pp. 225–33. *ScienceDirect*, <https://doi.org/10.1016/j.jad.2016.06.065>.
- Tsaneva, Stefani, et al. "Knowledge Graph Validation by Integrating LLMs and Human-in-the-Loop." *Inf. Process. Manag.*, vol. 62, no. 5, Sept. 2025, p. 104145.
- Tsaris, Aristeidis, et al. "Scaling Resolution of Gigapixel Whole Slide Images Using Spatial Decomposition on Convolutional Neural Networks." *Proceedings of the Platform for Advanced Scientific Computing Conference [New York, NY, USA], PASC '23, 2023*, pp. 1–11. *ACM Digital Library*, <https://doi.org/10.1145/3592979.3593401>.
- Valtolina, Stefano, and Daniele Fratus. "Local Government Websites Accessibility: Evaluation and Finding from Italy." *Digit. Gov.: Res. Pract.*, vol. 3, no. 3, Oct. 2022, p. 17:1-17:16. *ACM Digital Library*, <https://doi.org/10.1145/3528380>.
- Varoufakis, Yanis. *Technofeudalism - What Killed Capitalism*. VINTAGE, 2023. *Internet Archive*, <http://archive.org/details/technofeudalism-what-killed-capitalism-2023-yanis-varoufakis>.
- Vassilev, Apostol. *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*. NIST AI 100-2e2025, National Institute of Standards and Technology, 2025, p. NIST AI 100-2e2025. *DOI.org (Crossref)*, <https://doi.org/10.6028/NIST.AI.100-2e2025>.
- Veale, Michael, and Frederik Zuiderveen Borgesius. "Demystifying the Draft EU Artificial Intelligence Act — Analysing the Good, the Bad, and the Unclear Elements of the Proposed Approach." *Computer Law Review International*, vol. 22, no. 4, Aug. 2021, pp. 97–112. www.degruyterbrill.com, <https://doi.org/10.9785/cr-2021-220402>.
- Vettoretti, Martina, et al. "Advanced Diabetes Management Using Artificial Intelligence and Continuous Glucose Monitoring Sensors." *Sensors*, vol. 20, no. 14, Jan. 2020, p. 3870. www.mdpi.com, <https://doi.org/10.3390/s20143870>.
- Voit, Michael Matthias, and Heiko Paulheim. *Bias in Knowledge Graphs – an Empirical Study with Movie Recommendation and Different Language Editions of DBpedia*. May 2021.



Wachter, Sandra, et al. "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR." *arXiv:1711.00399*, arXiv, 21 Mar. 2018. *arXiv.org*, <https://doi.org/10.48550/arXiv.1711.00399>.

Wallace, Eric, et al. "Concealed Data Poisoning Attacks on NLP Models." *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, edited by Kristina Toutanova et al., Association for Computational Linguistics, 2021, pp. 139–50. *ACLWeb*, <https://doi.org/10.18653/v1/2021.naacl-main.13>.

Wang, Jing, and Xiuping Liu. "Medical Image Recognition and Segmentation of Pathological Slices of Gastric Cancer Based on Deeplab V3+ Neural Network." *Computer Methods and Programs in Biomedicine*, vol. 207, Aug. 2021, p. 106210. *ScienceDirect*, <https://doi.org/10.1016/j.cmpb.2021.106210>.

Wang, Qiang, et al. "Integrating Artificial Intelligence in Energy Transition: A Comprehensive Review." *Energy Strategy Reviews*, vol. 57, Jan. 2025, p. 101600. *ScienceDirect*, <https://doi.org/10.1016/j.esr.2024.101600>.

Wang, Xuesong, and Chuan Xu. "Driver Drowsiness Detection Based on Non-Intrusive Metrics Considering Individual Specifics." *Accident; Analysis and Prevention*, vol. 95, no. Pt B, Oct. 2016, pp. 350–57. *PubMed*, <https://doi.org/10.1016/j.aap.2015.09.002>.

Wang, Zheer. "Advancements and Challenges of Large Language Model-Based Code Generation and Completion." *Proceedings of the 1st International Conference on Modern Logistics and Supply Chain Management* [Singapore, Singapore], 2024, pp. 208–13. *DOI.org (Crossref)*, <https://doi.org/10.5220/0013271800004558>.

Wani, Ankit, et al. *Ethics in the Driver's Seat: Unravelling the Ethical Dilemmas of AI in Autonomous Driving*. [Detroit, Michigan, United States], Apr. 2024, pp. 2024-01–2023. WCX SAE World Congress Experience, *Semantic Scholar*, <https://doi.org/10.4271/2024-01-2023>.

Weidinger, Laura, et al. "Ethical and Social Risks of Harm from Language Models." *arXiv:2112.04359*, arXiv, 8 Dec. 2021. *arXiv.org*, <https://doi.org/10.48550/arXiv.2112.04359>.

Wu, Felix, et al. "Wav2Seq: Pre-Training Speech-to-Text Encoder-Decoder Models Using Pseudo Languages." *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. *IEEE Xplore*, <https://doi.org/10.1109/ICASSP49357.2023.10096988>.

Yaman, Albadawi, et al. *Real-Time Machine Learning-Based Driver Drowsiness Detection Using Visual Features*. <https://www.mdpi.com/2313-433X/9/5/91>. Accessed 26 May 2025.

Yang, Dapeng, et al. "Hybrid Physical Education Teaching and Curriculum Design Based on a Voice Interactive Artificial Intelligence Educational Robot." *Sustainability*, vol. 12, no. 19, Jan. 2020, p. 8000. *www.mdpi.com*, <https://doi.org/10.3390/su12198000>.

Zhang, Qin, et al. "The Rise of Small Language Models." *IEEE Intelligent Systems*, vol. 40, no. 1, Jan. 2025, pp. 30–37. *DOI.org (Crossref)*, <https://doi.org/10.1109/MIS.2024.3517792>.

Zhang, Ziyao, et al. "LLM Hallucinations in Practical Code Generation: Phenomena, Mechanism, and Mitigation." *Proc. ACM Softw. Eng.*, vol. 2, no. ISSTA, June 2025, p. ISSTA022:481-ISSTA022:503. *ACM Digital Library*, <https://doi.org/10.1145/3728894>.



Zhao, Xia, et al. "A Review of Convolutional Neural Networks in Computer Vision." *Artificial Intelligence Review*, vol. 57, no. 4, Mar. 2024, p. 99. *Springer Link*, <https://doi.org/10.1007/s10462-024-10721-6>.

Zhu, Beibei, et al. "A Survey: Knowledge Graph Entity Alignment Research Based on Graph Embedding." *Artif. Intell. Rev.*, vol. 57, no. 9, Aug. 2024.

Zwitter, Andrej, et al. "General-Purpose AI Regulation and the European Union AI Act." SSRN Scholarly Paper no. 4916400, Social Science Research Network, 31 July 2024. *papers.ssrn.com*, <https://doi.org/10.2139/ssrn.4916400>.