



Insights Into Incorporating Trustworthiness and Ethics in AI Systems With Explainable AI

Meghana Kshirsagar, University of Limerick, Ireland*

 <https://orcid.org/0000-0002-8182-2465>


Krishn Kumar Gupta, Technological University of the Shannon, Athlone, Ireland

 <https://orcid.org/0000-0002-1612-5102>

Gauri Vaidya, University of Limerick, Ireland

Conor Ryan, University of Limerick, Ireland

Joseph P. Sullivan, Technological University of the Shannon, Athlone, Ireland

 <https://orcid.org/0000-0003-0010-3715>

Vivek Kshirsagar, Government Engineering College, Aurangabad, India

ABSTRACT

Over the past seven decades since the advent of artificial intelligence (AI) technology, researchers have demonstrated and deployed systems incorporating AI in various domains. The absence of model explainability in critical systems such as medical AI and credit risk assessment among others has led to neglect of key ethical and professional principles which can cause considerable harm. With explainability methods, developers can check their models beyond mere performance and identify errors. This leads to increased efficiency in time and reduces development costs. The article summarizes that steering the traditional AI systems toward responsible AI engineering can address concerns raised in the deployment of AI systems and mitigate them by incorporating explainable AI methods. Finally, the article concludes with the societal benefits of the futuristic AI systems and the market shares for revenue generation possible through the deployment of trustworthy and ethical AI systems.

KEYWORDS

Agriculture, Black-Box Approach, Decision Support Systems, Deep Learning, Healthcare, Lime, Machine Learning, PDP, Responsible Artificial Intelligence, Security, Smart City

INTRODUCTION

Artificial Intelligence (AI) has extended into a significant technological shift in the recent decades such that each industry has been empowered in increased productivity, intelligent solutions, automation,

DOI: 10.4018/IJNCR.310006

*Corresponding Author

optimization, etc. to name a few. With the introduction of the determinant of trust, AI can no longer be treated as a black-box model without a clear understanding of what is going on inside. As AI ethics pose the single largest challenge towards widespread deployment, a trustworthy AI framework can help companies to design, develop, and deploy AI systems that they can trust. Better policies to manage ownership of personal data through adopting regulatory like the General Data Protection Regulation (GDPR), one can easily overcome the inappropriate usage of data, possible from uncovering behavioural patterns through data mining. AI systems can indeed be made more trustworthy and responsible by making them more traceable and explainable for the prediction of outcomes and decisions. In this research work, we present an overview of the determinants of AI – trust, the recent market trends and factors that can lead to responsible AI and software engineering.

LITERATURE METHODOLOGY

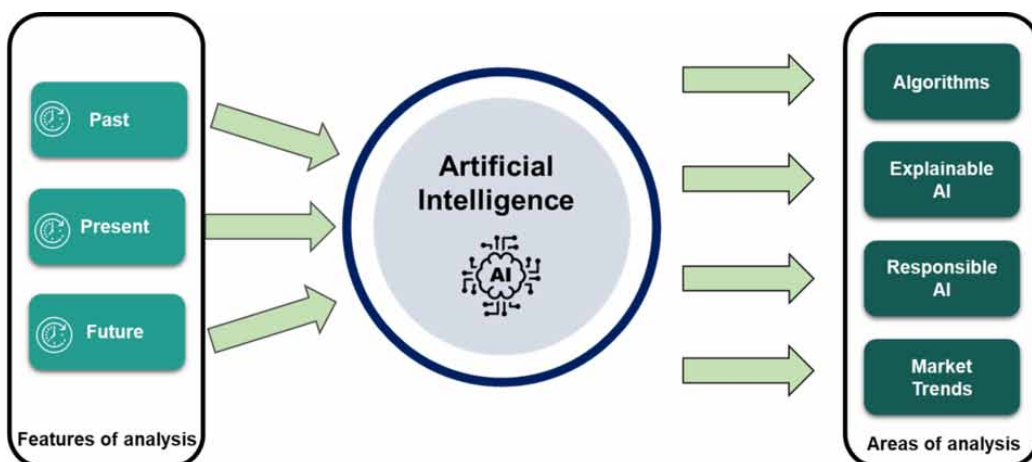
Our literature review is built by analyzing information across a range of sources. Figure 1 illustrates the pipeline of the proposed study where we integrate data based on information from the past, the current, and the predicted future for AI-powered applications. The objective of the study is to discuss the impact of AI-powered products on the wider community.

We present the evolution of AI, the Machine Learning (ML) algorithms in practice, applications in use, and the future of businesses and technologies with the integration of AI and its determinants like trust, big data and ubiquitous computing. We conducted a detailed study of all the ML and Deep Learning (DL) algorithms along with their use cases. We have an in-depth discussion on how incorporating explainability and interpretability into AI applications can lead to robust, trustworthy, fair and transparent AI systems. Finally, we bring to attention the importance of responsible AI engineering leading to regulated and accountable AI systems of the future.

The unique contributions of our proposed study are:

1. The evolution of AI and deep learning technology over the past seven decades;
2. Popular ML algorithms along with use cases drawn from diverse application domains;
3. Intelligent business models and market trends for industrial AI-powered products;
4. Incorporating Responsible AI for Trustworthy AI systems.

Figure 1. Pipeline of the proposed study



BACKGROUND

This section is split across a number of subsections starting with the roadmap of seven decades in AI, followed by a brief review on ML/DL algorithms, and concluding with diverse application domains of AI applications.

Evolution of AI

The term AI was first discussed in a context when a study was carried out jointly by three universities-Dartmouth, Harvard and Bell telephone laboratories. Soon followed the first programming language Lisp (Weinreb & Moon, 1981) for the AI researcher community.

The ML paradigm was one of its kind which learned from experiences like a human and could solve problems by manipulating sentences in formal languages. The early 60s witnessed the first-ever industrial robot 'Unimate' working in general motor's assembly lines. The mid-sixties gave birth to the chatbot 'ELIZA' (Weizenbaum, 1966) which facilitated dialogues between machines and humans. The late seventies led to the first autonomous vehicle 'Stanford cart' (Moravec, 1990). With the commencement of the 80s, AI capacity at automation was explored with the Xcon program (Barker et al., 1989). The mid-eighties saw the first MercedesBenz driverless car running on the empty streets of Munich. In the twenty-first century, the introduction of recommender systems (Chaudhari et al., 2018) designed to assist users in product selection choices based on their requirements led to the acceleration of the ecommerce market. IBM Watson (High, 2012), a real-time question answer-based system (Nagori et al., 2018) was a technological revolutionary breakthrough in the last decade. Figure 2 illustrates all these key events in the evolution of AI over the period of seven decades.

The exponentially growing data blending with the advancement in digital technologies has led to the surge of accelerated research in the domain of AI and ML techniques. The 2021 AI Index Report from Stanford explains that research interests in the domain of AI have sharply risen since 2016 as presented in Figure 3. At present, a lot of features and security tools have incorporated AI. AI has well-known applications in cyberattacks, medical devices, transportation systems etc. Falcon Platform, a security software by CrowdStrike uses anomaly detection for endpoint security using AI to defend against modern ransomware. Tessian, a software company in Landon uses ML technology to stop the security threats and data breaches caused due to human-digital interactions like business email compromise, accidental data loss and phishing attacks, within the enterprise. Business Intelligence (BI) applications are increasingly adopting ML in analytics processes. An example is SAP with HANA cloud platform used by Walmart for its high-volume sales transactions and customer information. Avanade, a management consulting company, uses AI for business analytics and predictive data visualizations. Pacific Specialty, an insurance company, uses a deep analytics platform to provide a better understanding to its staff about their customer base and policy trends. AI has been in use

Figure 2. Evolution of AI across seven decades (Tobin et al., 2020)

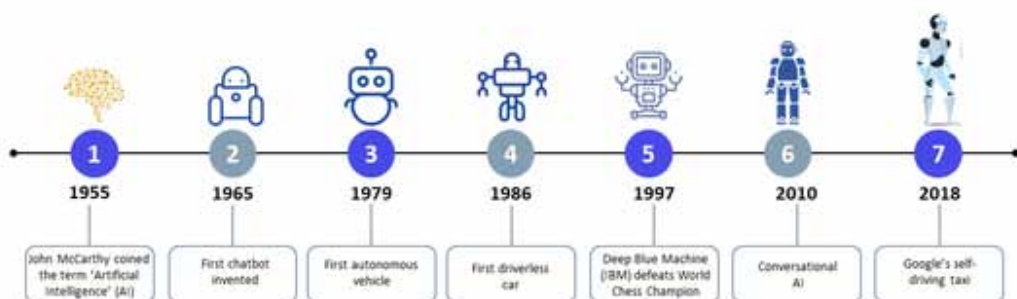
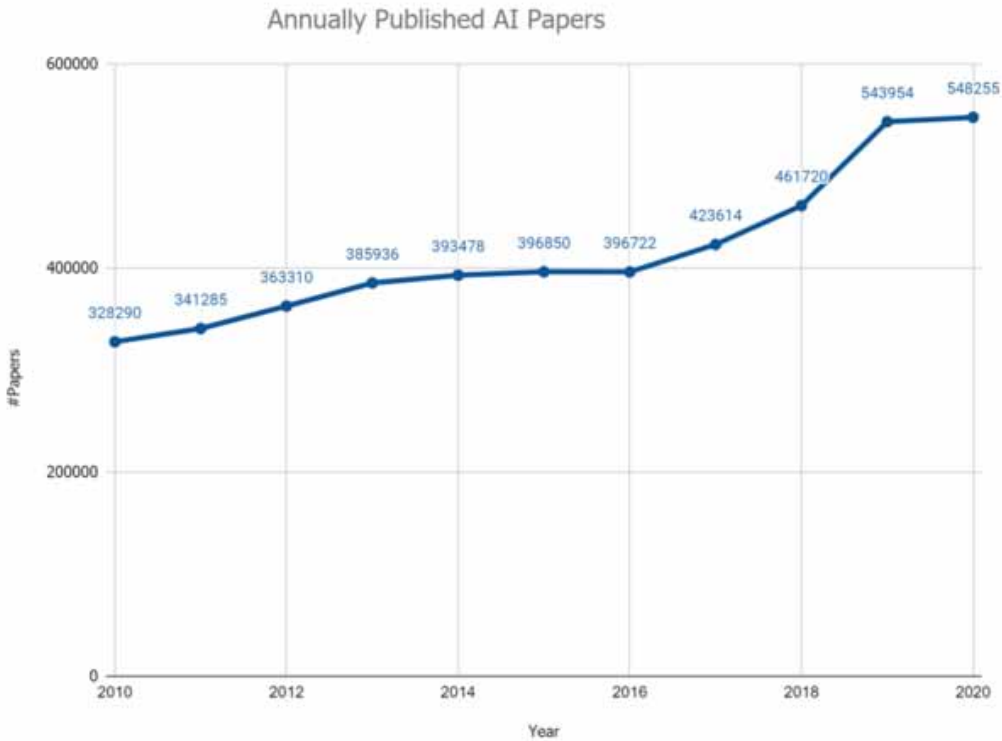


Figure 3. Number of annually published research papers in AI (Zhang et al., 2021)



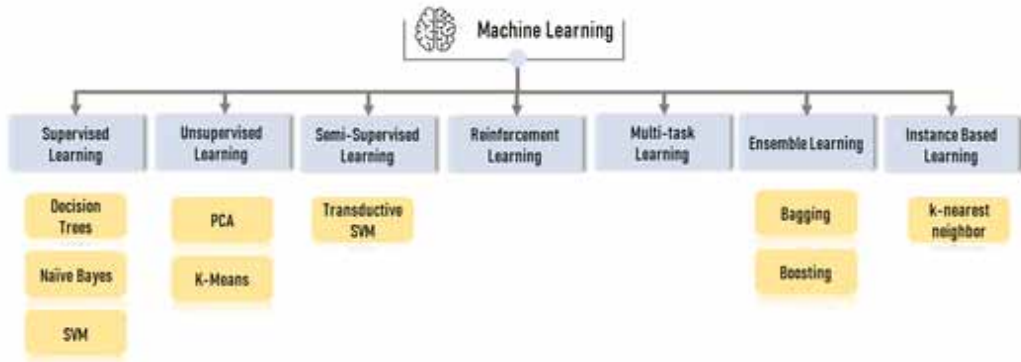
for security screening and crime forecasting and prevention. CompStat (COMPUter STATistics, or COMPARative STATistics) is a tool for predicting future crime hotspots. Interpol in Europe uses AI to fight child sexual abuse through the International Child Sexual Exploitation Image Database (ICSE DB) (Groff & LaVigne, 2002). In the healthcare industry AI/ML has been largely used in feature selection for large cancer datasets (Završnik, 2020), and data mining algorithms for Detection and Prediction of Disease class (Chitode & Nagori, 2013; Nagori et al., 2013), etc.

Convergence of AI/ML and Deep Learning

AI, a broader area, has as its subset ML and DL been a part of the subset of ML, fitting inside both. DL is driving today's AI explosion due to more complex inputs and outputs. ML is a subdomain under the AI umbrella that relies on data to teach a machine how to solve a problem. Speaking in general terms, ML techniques teach a machine to learn from their experience and improve their performance with respect to their previous mistakes. ML techniques are divided into three broad categories, supervised, unsupervised and semi-supervised learning. However, several other approaches divide these techniques further into seven different types as shown in Figure 4.

Supervised learning techniques train the machines to learn from already known patterns. The machine learns the pattern between the input and the output data. Decision tree classifiers, Support Vector Machines and Naïve Bayes are the three approaches under supervised learning techniques used for classification or regression problems like image classification from labelled datasets or predicting stock trends in the industries. Unsupervised ML techniques train the machines to learn the patterns from the data on their own to decipher unknown trends from the data. The Principal Component

Figure 4. Types of Machine Learning Techniques (Batta, 2020)



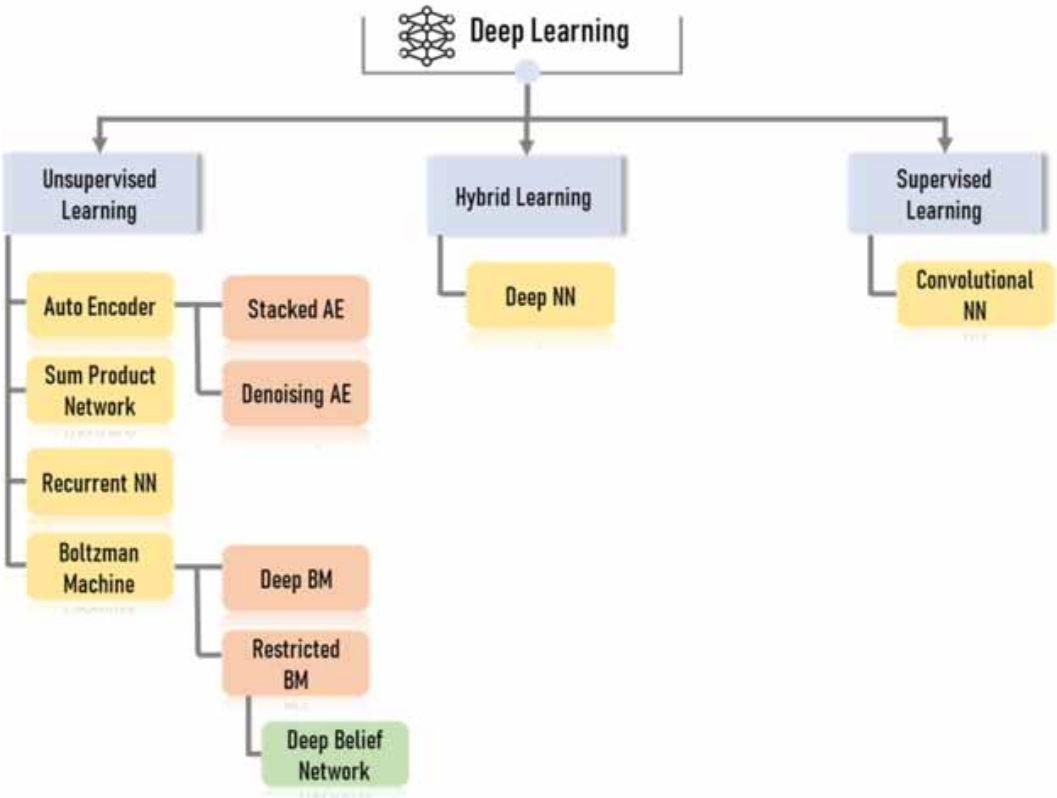
Analysis technique is used for reducing redundant features while K-means is used for clustering. These techniques are used for feature reductions or clustering problems in domains like developing marketing strategies based on customer experience, or clustering from a vast number of samples. Semi-supervised learning falls in between supervised and unsupervised learning techniques. Speech analysis is a well-known example of a semi-supervised technique as the machine initially learns from the available data and continuous learning by time according to the user inputs. Reinforcement learning paradigm trains the machine in such a way that the machine behaves ideally in a given scenario to maximize the cumulative reward. Agent-based systems automate intelligent decision-making systems. Multi-task learning trains the machine to use the knowledge in one task to be reused in another leading to better efficiency. Bagging and boosting are two variants under ensemble techniques. These techniques adjust the variance and bias of the machine to boost performance. Instance-based learning is the paradigm that adapts to the new environment and new data by learning to relate it with training data. K-nearest neighbor is the variant of instance-based learning and is used in applications such as personalized recommender systems (Bindra et al., 2021).

DL is a subdomain under AI that allows the model to learn patterns from the data itself with multiple levels of abstraction (Pouyanfar et al., 2019). Deep learning techniques are divided into three categories, unsupervised learning, supervised learning and hybrid learning as illustrated in Figure 5. Convolutional Neural Networks (CNNs) are inspired by the human brain neuron's structure to extract from the two dimensions of the data useful in applications like speech recognition, computer vision and Natural language Processing (NLP) (Hubel & Wiesel, 1962). Recent state-of-the-art deep learning systems significantly highlight the importance of transfer and unsupervised learning. CNNs have extended to bridge the gap between supervised and unsupervised learning to generate large synthetic datasets with Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) and Deep convolutional generative adversarial networks (DCGANs) for unsupervised learning (Radford et al., 2015). Recurrent Neural Networks (RNN) use sequential information such as sequences of words that lead to meaningful knowledge. The three techniques are intertwined with each other and overlap in certain cases.

Future with AI

AI is evolving progressively from intelligent assistants like Alexa and Siri to object detection from real-world traffic data (Kshirsagar, More, Lahoti, et al., 2022; Vaidya et al., 2022) to train self-driving cars. We are already surrounded by AI applications starting with smart devices – phones, wearables, voice assistants, etc. AI applications are projected to have a lasting impact on every industry imaginable

Figure 5. Classification of deep learning techniques



(Grace et al., 2018) in the foreseeable future. Be it precision medicine or climate change or tracking hate speeches to intelligent surveillance systems, AI solutions will revolutionize society.

Many researchers in the field of AI explore the use of AI to create another AI. Some examples include Google’s AutoML (*Cloud AutoML Custom Machine Learning Models* | Google Cloud, n.d.) project that employs machine learning expertise to generate other high-quality machine learning models in minutes. AI is not limited to science, technology, healthcare, etc. but it also finds use cases in the field of music and art. For example, initiatives like Google Magenta (*Magenta*, n.d.) are investigating the possible contributions of AI in producing compelling art and music. We present below a few of the domains where AI solutions have been proved useful.

Customer Services

Google is working on AI assistants like Google Duplex that can place human-like calls and have natural conversations to make appointments at, say, a hotel, restaurant, or a neighborhood hair salon.

Education

AI is used to digitize textbooks and is used as an educational tool to guide students by evaluating different mock tests and providing them personalized feedback. The use of AI is highly explored in higher education for several activities like plagiarism detection and academic research etc.

Manufacturing

AI-enabled machines are empowered to collect and process data, recognize patterns, and learn and adapt according to the dynamic work environment to facilitate productivity. AI-powered robots also perform tasks like stacking products and help run the equipment smoothly.

Media

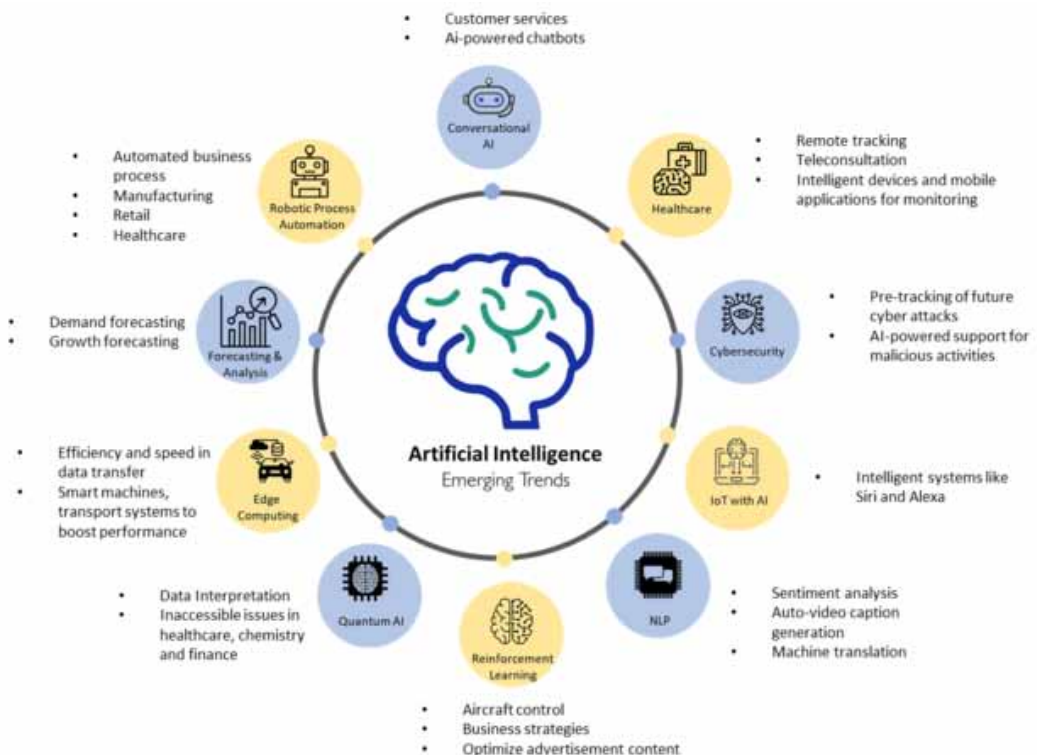
Companies like Bloomberg uses Cyborg technology to analyze and make quick sense of complex financial reports. Journalism is harnessing AI for many applications like identifying fake news and generating flagging alerts as soon as a trend or anomaly emerges.

Transportation

AI has the potential to ease traffic congestion, make it more efficient, and hence reduce fuel consumption and improve air quality (Kshirsagar, Vaidya, Rajguru, et al., 2022; Niestadt et al., 2019). Although it could take a decade or more, autonomous cars will one day ferry us from place to place.

Figure 6 illustrates several emerging application areas for AI with security, robotics and healthcare at the forefront. The goal of AI is to design 'general' intelligence for performing unpredictable tasks rather than 'specific' intelligence, which can perform only a single task outstripping human skills. The learnings through direct interaction of the object or the model to the physical environment are important for domains like robotics and driverless cars for understanding the cause-effect behavior. The upcoming AI models will be designed based on the pre-trained models and will be implemented on the energy-efficient hardware (López de Mántaras, 2018). Although the increasing importance and

Figure 6. Emerging trends for AI (Johns, 2020)



impact of this upcoming technology, its use depends on the use case of the target application. Some systems can be developed with strong mathematical foundations and theoretical concepts but never tested while there may be others developed through black-box approaches with extensive testing. In such cases, the second AI system will be more preferable for deployment as compared to the first. Furthermore, blockchain will facilitate AI scale and lead to data sharing frameworks that are secured thus facilitating in creation of trustworthy and transparent business models (Kshirsagar, Vaidya, Yao, et al., 2022; Nagori et al., 2020; Qazi et al., 2020; Yao et al., 2021).

ADVENT OF XAI

The increasing impact of AI on traditional use cases also increased the concerns for trust and explainability that led to the need for the processes that could explain the results of the models. In particular, decision-making models like credit card risk assessment, health predictions, etc. need an explainable mechanism (Figure 7) to check the possibility of bias or error in the result. Around forty years ago, this concept started gaining importance which has covered a lot of domains to date.

There are many real-world examples where AI has gone awry due to systems discriminating against people based on gender, race, age. These include criminal sentencing decisions biased due to race such as blacks, online recruitment systems biased towards preferences towards the male gender, or financial institutions screening mortgage applications on basis of age, gender and race. As AI systems continue to be deployed, such biases can result in serious societal consequences leading to angry customers. From a business perspective, it can potentially damage companies leading to lawsuits, regulatory fines etc. Trustworthy AI systems can help towards alleviating these risks by having an ethical framework in place to articulate trust and help towards making fair decisions by reducing discriminatory bias.

Features of XAI

The AI models should be **transparent** in representation to verify the results and the intermediate judgements. The data, features and AI models should be presented to enable users to understand what is happening inside the AI model.

The seven pillars as shown in figure 8 lay the foundation of the XAI. The explainable data should be in context to the **domain** of the system that aligns the knowledge and data together. The explanation of the unpacked model should be **consistent** across multiple ML models and multiple runs. Multiple runs and models shall deliver the same context of the results produced. The explanation of the data should be clear and precise in context to technical terms. The explanation should be in such a format

Figure 7. Incorporating xAI with traditional AI

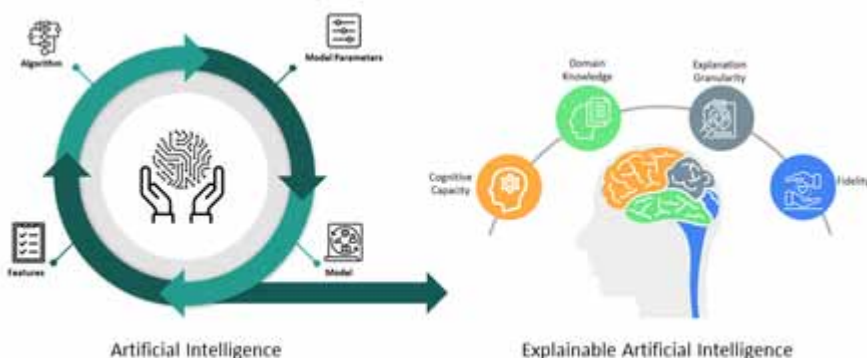
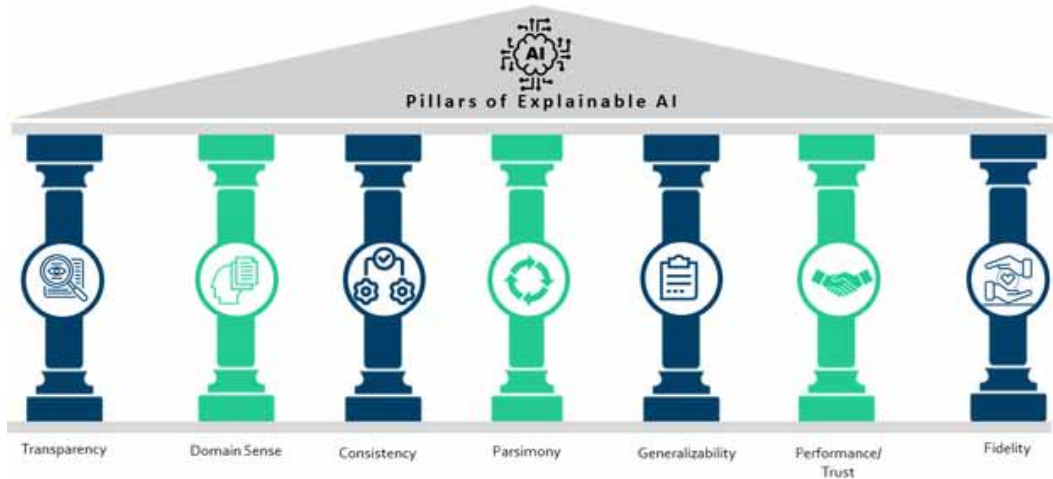


Figure 8. Pillars of xAI



that it can be used for a broader group of situations leading to *the generalizability* of the model. The results of the models should be able to be measured in terms of *performance* leading to *trust* among the users and model results. The explanation and the results of the model should align with each other leading to fidelity among users and system providers. Following these pillars, while designing an xAI model, leads to a trustable and ethical model, boosting the performance and capabilities of the model in the domain.

XAI Methods and Algorithms

Some models are designed in a way that they are self-explainable or ad-hoc with a little trade-off between accuracy and interpretability of the features. However, these techniques are complex to implement and the literature mostly shows the use of post-hoc methods or interpretation after the classification. These methods include Respond-CAM (Zhao et al., 2018), saliency maps (Kim et al., 2019; Simonyan et al., 2014), SHAP, and Grad-CAM. Using attributes as an interpretability factor, there are two sub-methods: perturbation-based approaches and back-propagation-based approaches. The example of the former one is SHAP where the attribute value of each feature is extracted by calculating the impact on the final output. The latter one takes the whole data as input and calculates the impact in either a backward pass or a forward pass. Examples of using these methods in the literature include (Lee et al., 2018), where authors presented a comparative analysis of PG-CAM to Grad-CAM for interpretability of MRI images in brain cancer prediction models. Authors (Hamm et al., 2019) used the CNN model for classifying lesions in the liver for cancer.

Model-Agnostic Interpretability Methods

Several model-agnostic interpretabilities are helpful for researchers in interpreting complex ML models. Typically, many machine learning models are evaluated and used together to solve a task, thus it becomes easier to use model-agnostic interpretability methods as the same method can be used for any type of ML model. Different model-agnostic interpretability methods are discussed below.

Partial Dependence Plot (PDP)

The PDP shows the change in the average predicted value of a machine learning model as a specified feature varies over their marginal distribution (Friedman, 2001). It can show whether the relationship between the feature and target is linear, monotonic, or more complex. Green and Kern (Green & Kern, 2021) in their voter mobilization experiment used PDP to understand the relationship between the predictors and the conditional average treatment effect. Under asymmetric classification costs, Berk and Bleich (Berk & Bleich, 2013) used PDP to model predictor-response relationships. Since PDP shows the average marginal effects, the heterogeneous effects can get hidden (Molnar, 2019).

Individual Conditional Expectation (ICE)

ICE plots disaggregate the output of classical PDPs. ICE helps explain the change in the predictions of the model as a particular feature varies (Goldstein et al., 2015). Unlike PDP, the ICE curves can uncover heterogeneous relationships and are more intuitive to understand than PDP plots. Since it can only plot one feature meaningfully as two features need drawing of several overlapping surfaces, it sometimes can be overcrowded. Also, it might not be easy to see the average effect as is possible with PDP. In a recent work of severity prediction of Covid-19, H. Wu et al used PDP and ICE to visualise the relationship between C-Reactive Protein (CRP) and N-Terminal pro-Brain Natriuretic Peptide (NTproBNP) (Wu et al., 2021). Elshawi et al used model-agnostic methods to explain their ML model designed for predicting the risk of hypertension based on the cardiorespiratory fitness data (Elshawi et al., 2019).

Permutation Feature Importance

With the Permutation Feature Importance (PFI), the effect of the selected feature on the overall predictive power of the model is tried to be understood considering the decrease in the model score when a single feature value is randomly shuffled (Breiman, 2001), which breaks the relationship between the feature and the true outcome. This is especially useful to examine the sensibility of non-linear or opaque estimators. (Galkin et al., 2020) in their work on microbiological influence on age prediction used PFI to assess which taxa abundances play the greatest role in the microbiological age prediction while examining the relationship between human gut taxonomic profiles and chronological age. PFI provides a highly compressed, global insight into the model's behavior.

Global Surrogate

The global surrogate method is an interpretable model that is trained to approximate the prediction of a black-box model. The global surrogate model method is flexible. Recently et al presented a global surrogate-assisted covariance matrix adaptation evolution strategy (CMA-ES) by implementing a global linear-quadratic surrogate model as an add-on module to the *pycma*, a Python package of CMA-ES (Hansen, 2019).

Local Surrogate (LIME)

Local Interpretable Model-agnostic Explanations (LIME) or Local Surrogate models are used to train interpretable models to approximate the individual predictions of the black-box ML models (Ribeiro et al., 2016). Unlike global surrogate models, it tries to explain the individual predictions rather than the whole model. A theoretical analysis of LIME is presented in (Garreau & Luxburg, 2020).

TRADITIONAL AI/ML USE CASES

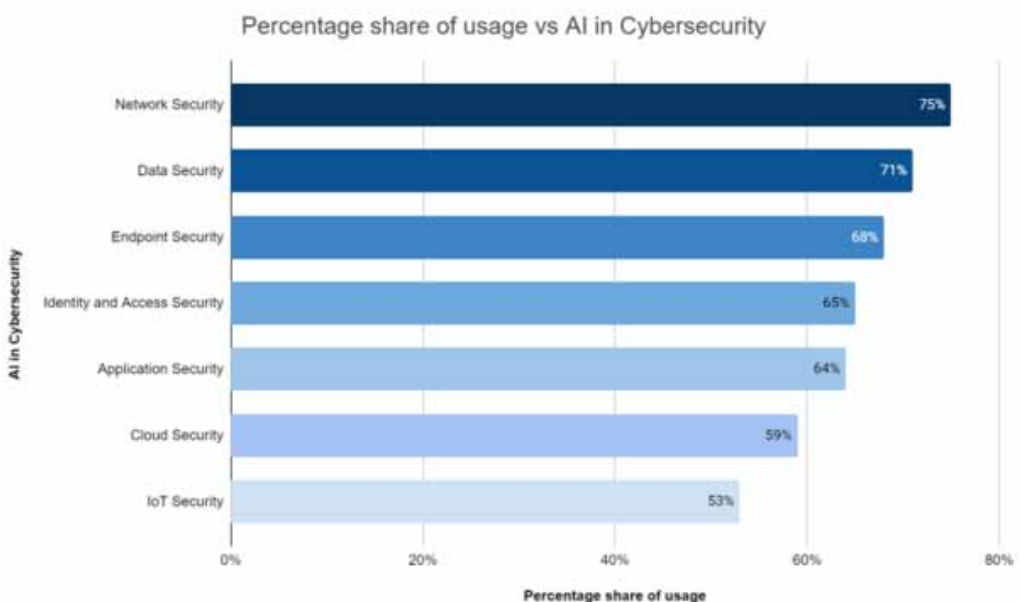
This section will discuss some of the applications of AI and ML in the significant domains of security, agriculture, and smart cities.

AI/ML in Security

AI models such as Deep Neural Networks (DNN) were implemented as smart cyber security models for threat and intrusion detection. As per a report in (Tolido et al., 2019), 69% of companies agree that it would not be possible for them to respond to cyber-attacks without AI. The Insider Threat Report (2018) (*Insider Threat Report*, 2017) found that 90% of organizations are vulnerable to insider attacks. Another report published in 2018 (*Threat Hunting Report*, 2018) found that the possibility of threat occurrence has doubled in past years making advanced threat detection to be a top challenge for 55% of Security Operation Centers (SOC). AI follows the three paradigms of Learn, Reason, and Augment (*Beyond the Hype: AI in Your SOC*, 2020). It learns through getting trained on massive data from both structured and unstructured data and improves its knowledge and understanding of cyber security threats using ML and DL. The insights are used to identify the relationships between malicious and suspicious agents that help in providing a curated analysis of threats (Figure 9).

Common AI/ML-based cybersecurity solutions include k-nearest-neighbour, support vector machine, neural network, decision tree, Bayesian network, etc (V. Kshirsagar & Joshi, 2015). The various algorithm selection depends on the type of problem domain. An AI-based data security tool can reduce the risk either by automation and detection or by providing enhanced capabilities to the SOCs to launch an orchestrated response. Few of the AI-enabled data security tools are Security Information and Event Management (SIEM) and User and Entity Behavior Analytics (UEBA) and Cylance by Blackberry. A pioneer cloud-based identity and access management vendor called OneLogin has provided IAM security tools that use AI to provide high-level security. To explain anomalous behaviors, simMachines uses similarity-based explainable AI (XAI) technology. A surge in the IoT's user base, network bandwidth, data privacy, demand on-device data processing, created a need for more reliable IoT security solutions to mitigate the risk against the data at rest, in use and in motion. The limitations of the traditional rule-based security infrastructure (V. Kshirsagar & Joshi, 2016) have led to the use of AI as the primary workforce in cyber security.

Figure 9. AI-driven domains in cybersecurity. Top AI use cases for Cyber Security in Organisations. (Selected countries: Australia, France, Germany, India, Italy, Netherland, Spain, Sweden, United States, 2019, 850 Respondents), Source: (Marin, 2005)



AI/ML in Smart Cities

The voluminous rise in traffic flow has led to issues like traffic congestion, pollution, accidents, etc. Injecting some level of automation using AI can overcome the incompetence of conventional traffic management systems. Several popular object detection algorithms such as regional convolutional neural network (RCNN), Mask R-CNN (He et al., 2017), Faster R-CNN (Ren et al., 2017) have been used in intelligent transportation systems (M. Kshirsagar et al., 2021). Another object detection algorithm, You Only Look Once (YOLO) (Bochkovskiy et al., 2020), has been utilised far and beyond as a state-of-the-art in intelligent transportation systems. YOLO and deep convolutional neural networks (DCNN) have been implemented to detect traffic congestion from camera images with an accuracy of 90.2% and 91.5% respectively (Chakraborty et al., 2018). Another object detection algorithm CenterNet (Liu et al., 2020) has been popular in real-time image detection due to its shorter training time and faster inference. Mandal et. al (Mandal et al., 2020) developed an automatic traffic monitoring system using Mask R-CNN, Faster R-CNN, YOLO, and CenterNet. Intelligent traffic monitoring systems help to identify traffic violations, crossing lanes and can help in environmental sustainability by reducing traffic congestion. AI is extensively being used in multiple types of research on driving unmanned vehicles or computer-controlled cars, driver behavior modelling, forecasting traffic streaming, etc.

AI is having an enormous impact on the waste management and recycling sectors. Intelligent bins and smart sorting are two major applications where the role of AI has been pursued successfully. Through the incorporation of sensors and ML algorithms, different types of garbage are distinguished and sorted. Also, smart bins allow the users to locate a bin nearest to their location, thereby preventing bins from overflowing. In smart sorting, the items are placed on conveyor belts and scanned by cameras. These items are identified by DL algorithms and robotic arms are used to pull the items off the belt for further processing. AI/IoT tools can help towards the segregation of different plastics thus making them a reliable source to be used in the circular economy of plastic (Chidepatil et al., 2021). Classifying waste into biodegradable/ non-biodegradable classes help to identify its suitability for disposal or not and can facilitate choosing proper disposal techniques for different categories of waste (Nagori et al., 2019).

AI/ML in Agriculture

The growing population, demand for food and fibre consumption and climatic changes have created tremendous challenges in agriculture. According to the World Population Prospects 2019 (*How to Feed the World in 2050*, 2009), the population is projected to grow from 7.7 billion in 2019 to 9.7 billion in 2050, while the land under cultivation will account for 4%. The United Nations Food and Agriculture Organization (FAO) expects that the annual cereal production needs to rise to 3.1 billion tonnes and meat to be raised by 200 million tonnes (*How to Feed the World in 2050*, 2009). To feed such a large population, where 70% of the population will be urban, a smaller cultivational land and rural labour force will be a bottleneck in food production. The yield maximization in farming faces several challenges such as improper soil treatment, plant disease, pest infestation, climate and weather uncertainties and market price fluctuations etc. Dominant key factors like the monsoon, weather prediction and soil parameters (Agarwal et al., 2018) are major factors in predicting the yield (Guardo et al., 2018). Farmers in recent years are increasingly turning towards high tech solutions for improving yields and reducing costs. Large scale cultivations are taking advantage of using self-driving tractors and satellite imagery using artificial intelligence to increase their farming efficiency. ML with the help of sensors installed in the farms is facilitating better decision making about the type of fertilizers to be used and time of irrigation etc. (Courtois, 2019). Microsoft Corporation is using AI for rendering farming solutions to farmers in Hyderabad, India. In many villages, farmers are receiving automated updates on a call about their crop risk based on pest attacks, crop stage and weather conditions etc. The price forecast using AI helps in planning the Minimum Support Price (MSP). In collaboration with the International Crop Research Institute for Semi-Arid Tropics

(ICRISAT), Microsoft has developed an AI Sowing App that updates farmers about the optimal date to sow (*Digital Agriculture: Farmers in India Are Using AI to Increase Crop Yields*, 2017). A Pest Risk Prediction API developed by Microsoft, with United Phosphorus Ltd (UPL), uses AI and ML to predict the risk of pest attack. Other popular adoptions include smart harvesting by Deere & Co. which incorporates high-resolution cameras and sensors linked to AI systems. The system monitors the collection of grains to optimize how much grain is chopped from each stalk thus minimizing waste. AI in farm surveillance is another game-changer. Twenty20 Solutions, a security service company in Texas, is providing smart surveillance using AI and ML tools to monitor crop fields in real-time through video feeds and alerts the owner if any human or animal breaches are detected. Future farming practices will increasingly rely on AI/ML-driven intelligent decision support systems for controlling chemical sprays, yield optimization, price assessments, crop disease detection etc.

TRANSITION TOWARDS XAI APPLICATIONS FROM TRADITIONAL ML

Although DNNs have great generalization and prediction skills, they are opaque machine learning models. This can lead to creation of decision support systems that are not justifiable or legitimate. Therefore, there emerged a growing need of endowing ML models with explainability as xAI techniques can serve to verify and certify model outputs and enhance them with desirable notions such as trustworthiness, accountability, transparency and fairness. This section below highlights some of the use cases for AI systems that are powered by incorporating the explainability methods as explained above.

Clinical Decision Support Systems

Traditional black-box ML-based AI models lack explainability. xAI addresses some of the issues around the traditional black-box AI system by explaining and interpreting their diagnosis, predictions, and recommended actions to stakeholders (Garreau & Luxburg, 2020). xAI driven systems create more understandable, interpretable, and reliable models, thereby improving the quality of predictions (Shickel et al., 2019). The issues around lack of explainability, transparency, and human understanding of how AI works, are reasons why people have little trust in the AI healthcare system. To address the major medical AI challenge: explainability, from a Western legal point-of-view, the three core fields for explainability: (1) Informed consent, (2) Certification and approval as medical devices (acc. to Food and Drug Administration/FDA and Medical Device Regulation/MDR) and (3) Liability needs to be addressed. The interpretability methods vary according to data modality, locality, and specificity in medical images such as MRI, X-rays, brain, skin, retinal, breast, and CT scans. If the interpretability follows a single sample, the model generally uses the SHAP algorithm while if multiple samples or the entire model is used, Garson's Algorithm or Global Model Explanation Techniques (Garson, 1991) via Recursive Partitioning (Yang et al., 2018) are used. Under data modality techniques, Grad-CAM (Selvaraju et al., 2020) is used for MRI images and LIME (Ribeiro et al., 2016) technique is applicable for all types of image data types. Some algorithms are specific to ML and DL models, e.g., Grad-CAM provides an explainable feature for CNN but does not support LSTM models while SHAP is not constrained to a particular model or datatype. In healthcare, AI comes in the form of clinical decision support systems (CDSS), facilitating clinicians in the diagnosis of disease and planning treatments. AI-based CDSSs apply AI models trained on data from patients matching the use case at hand. According to the government, UK healthcare will be hugely driven with AI playing a key role towards transforming, preventing and in early diagnosis and treatment of chronic diseases. But doctors and clinicians remain sceptical about the AI healthcare system. Studies have revealed that among the 30% of clinicians' respondents lack trust in AI. A huge number of public correspondents in the UK, 61% are also unwilling to engage with AI for their healthcare needs. There needs to be two levels of explainability in healthcare. First-level explainability allows us to understand how the system arrives at conclusions in general. Second-level explainability allows us to identify which features

were important for an individual prediction. For e.g., the explanation module enabled the xAI-EWS to pinpoint which clinical parameters at a given point in time were relevant for a given prediction.

Another use case of xAI in healthcare is in reducing the features by extracting the significant ones to develop an explainable system (Porto et al., 2021). The authors followed the ranking algorithm for extracting the relevant features and tested on multiple ML algorithms to derive the best underlying model. They tested and proved with the real-time dataset of Statlog (Heart) data set from the University of California's Automated Learning Repository how extracting significant features reduced the cost, increased the interpretability, and boosted the performance of the predictive system. Interpretable ML was used to identify biomarkers to predict patient's antitumor response to Immune Checkpoint Blockers based on patients tumour RNA-SEQ DATA (Lapiente-Santana et al., 2021)

Explainability will provide answers in resolving the disagreement between an AI system and human experts. The results of explainability represented visually will allow the clinicians to examine the factors contributing towards a final recommendation. The clinicians using their experience can make an informed decision on whether to rely on the system's recommendation. This can pave the way towards strengthening their trust in the system. Explainability can support developers and clinicians to detect and correct biases such as discriminating against people for a medical AI system which can be a potential source of injustice, ideally at the early stage of AI development and validation, e.g., by identification of important features indicating a bias in the model.

Image Classification Systems and Face Recognition Systems

Another use case of using xAI in real-world systems is Google using xAI for image classification to understand how image models are deployed on Cloud AI. xAI uses two methods to deploy XAI for images: XRAI (Kapishnikov et al., 2019) and Integrated Gradients (IG) (Sundararajan et al., 2017). XRAI uses the color heatmap that denotes which region of the image impacted the most while IG returns the individual pixel that impacted the prediction. XRAI is widely used for images in nature as there is a wide range of different colored objects in the picture. This method gives an overall summary of the object in the picture like a boat or a dog. As IG uses individual pixels, it provides vast granularity. This feature of the IG method makes it suitable for medical and lab images. For a promising future with Virtual Reality and Augmented Reality, a significant milestone has been achieved with real-time face detection from videos with online learning (Thies et al., 2018b). With xAI, this research area has also extended towards a broader goal of bias free AI with applications detecting manipulations in visual media (Dolhansky et al., 2019).

Advanced Driver Assistance Systems

Advanced Driver Assistance Systems (ADAS) (Dickmanns & Mysliwetz, 1992) are the driver-assistance systems that aim at boosting the safety of the driver as well as the road. Recent studies in the literature (Lorente et al., 2021) focus on three factors in driving vehicles on the road: The distractions of the driver while driving (talking on the phone or texting), driver emotions (positive, negative, and neutral) that affect the driving style and some unavoidable environments (intentional or unintentional driving practices). The use of XAI in ADAS aims at giving justifiable interpretations about a driver's behavior at a particular moment. Such examples include which image triggered the emotions of the driver, or which environment or situation increased the speed of the vehicle, whether the behaviour was intentional or unintentional, and so on. These interpretations not only increase safety but also provide trustworthy information to insurance companies or law enforcement. Hence, the interoperations will also help in debugging the systems to improve their accuracy and reduce the bias if any.

RESPONSIBLE AI SYSTEMS FOR POSITIVE SOCIETAL IMPACT

The challenges of adopting AI systems in critical applications such as autonomous driving, to harnessing data from wearables for personalised healthcare has raised concerns around regulation and accountability. To tackle such issues AI systems, need to be designed and deployed with responsibility to empower businesses and society. This can be made possible only through initiatives taken by the leading AI giants such as Meta and Google to engender trust in AI systems by addressing privacy and security challenges in a transparent way. Hence, to ensure positive benefits by adopting AI to solve challenging problems such as the development of new drug designs to Precision Agriculture driven by AI to autonomous drones for monitoring fields for droughts and pests, the AI systems will be a lucrative option if the ML systems are developed responsibly complying to regulatory guidelines and policy standards.

AI technology can provide next-generation autonomous systems such as autonomous tractors, robotic picking of fruits and flowers, autonomous spraying and automatic milking. The impact of Artificial intelligence on society will be observed through AI-powered technologies facilitating reduction in greenhouse gas (GHG) emissions and the potential likelihood of leading to reduced GHG emissions by 16% by 2030. Data mining techniques will identify futuristic energy demands thus assisting the suppliers for optimizing the use of resources and filling in gaps with renewable resources while reducing waste. Hence, AI can power climate change strategies by aiming at reducing waste in all forms such as material, money and time and circularising economy models. AI-powered autonomous vehicles, including shared cars and smart transportation systems in some cities, will further help curtail emissions and reduce commuting times. Revolutions in patient-centric systems and powerful clinical decision support systems can lead to reduced mortality by providing better access to vaccines, lifesaving drugs, and less susceptible to diseases by better management of sewage and providing clean air, fresh water and optimizing food distribution to eliminate starvation and malnutrition. Personalized healthcare systems for recommending life's outcomes can assist individuals to manage healthy lifestyles and possibly reduce hospital visits.

MARKET SHARES FOR AI SYSTEMS

With the vast opportunities and applications, AI offers it is predicted that the global revenue market for AI-driven services and products will drastically increase from 10.1 billion dollars in 2018 to 126 billion dollars in 2025 as illustrated in figure 10 (*Revenues from the Artificial Intelligence (AI) Software Market Worldwide from 2018 to 2025*, 2020)

On the other hand, the global market of xAI is exponentially growing with an estimated size of USD 3.55 billion as of 2019 and is expected to be USD 21.78 billion by 2030 (*Explainable AI (XAI) Market*, 2021).

Industries and markets are evolving with business strategies like reducing costs, automation, innovation, and managing risk factors with advanced analytics of AI. This estimates that the markets will grow in the coming few years. As represented in figure 11 (*Machine Learning Market*, 2020), in 2017 in a survey by market research future, it was recorded that 44% of the global machine learning market was centred in North America, 29% in the Asia Pacific, 21% in Europe and 7% for the rest of the world based on the vertical, organization size and components of the industry.

CONCLUSIONS AND FUTURE SCOPE

This research article discussed the roadmap leading towards seven decades of a digital revolution powered by AI. Through a myriad of use cases in domains such as security, smart cities, agriculture and healthcare the authors make it evident that AI technology is quickly becoming a potential disruptor and enabler for all industries. The authors highlight some of the concerns inherent in

Figure 10. Forecasted market revenue generation from AI Software market

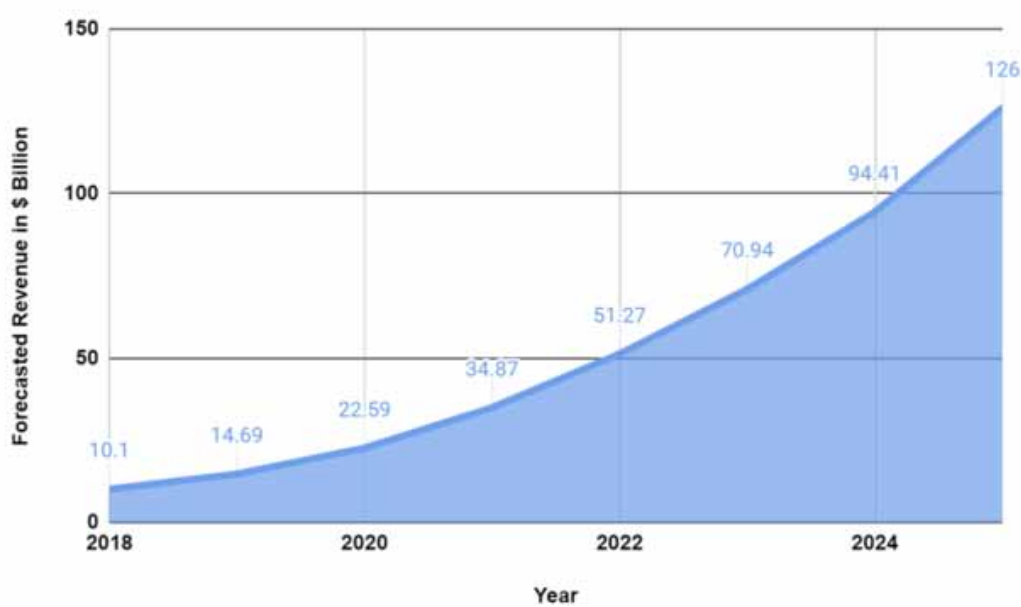
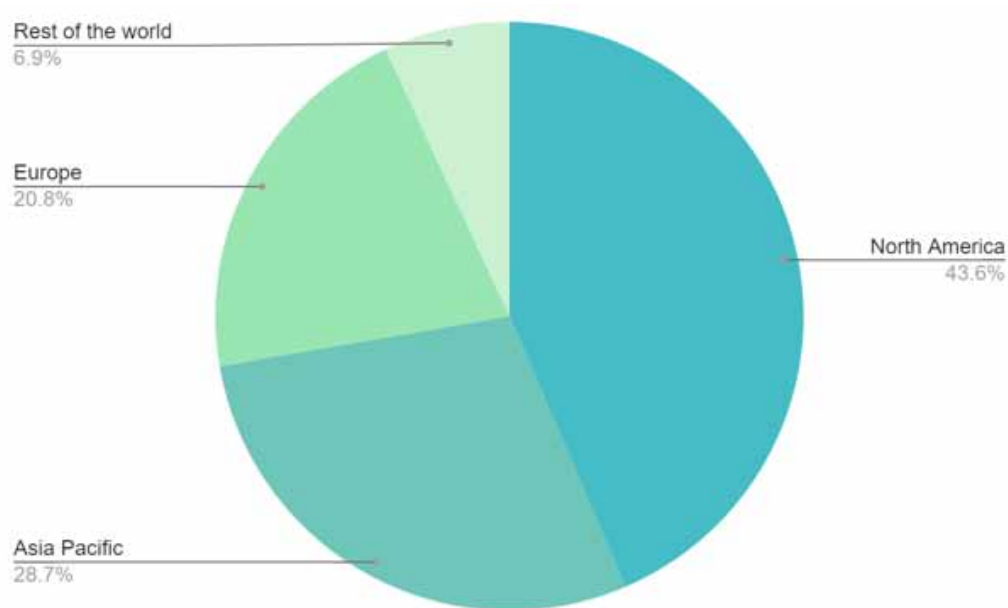


Figure 11. Global Machine Learning Market by region in the year 2017



existing AI systems leading towards hesitancy in its widespread adoption and possible mitigation through incorporating explainability methods. Explainability will assist developers to validate the performance of AI-driven systems based on enforcing privacy and security commitments to meet regulatory standards prescribed by different countries. Equally important would be to allow users

of the AI systems affected by a decision or an outcome to understand the workings by making the system fair and transparent. The authors depict that in future AI technologies can be leveraged for the betterment of society by supporting diversity and inclusion, preserving the natural environment by addressing climate change challenges, supporting the healthcare industry by assisting doctors in disease diagnosis, robotic surgery and precision medicine. Thus, AI can help communities in which humans inhabit through betterment of public services. The article concludes through a glimpse of the market economy in terms of possible revenues that can be harnessed by companies towards deploying trustworthy AI applications. The authors believe this comprehensive study will help the researchers in better understanding of the parameters of AI that can lead to responsible solutions leveraging the power of AI with xAI and ever advancing techniques.

ACKNOWLEDGMENT

Conflict of Interest

The authors of this publication declare there is no conflict of interest.

Funding Agency

This research was supported by the Science Foundation Ireland [grant number #16/IA/4605].

REFERENCES

- Agarwal, S., Bhangale, N., Dhanure, K., Gavhane, S., Chakkarwar, V. A., & Nagori, M. B. (2018, July). Application of Colorimetry to Determine Soil Fertility through Naive Bayes Classification Algorithm. *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. doi:10.1109/ICCCNT.2018.8494113
- Barker, V. E., O'Connor, D. E., Bachant, J., & Soloway, E. (1989). Expert systems for configuration at Digital: XCON and beyond. *Communications of the ACM*, 32(3), 298–318. Advance online publication. doi:10.1145/62065.62067
- Batta, M. (2020). Machine Learning Algorithms - A Review. *International Journal of Science and Research*, 9(1), 381.
- Berk, R. A., & Bleich, J. (2013). Overview of “Statistical Procedures for Forecasting Criminal Behavior: A Comparative Assessment.” *Criminology & Public Policy*, 12(3), 511. Advance online publication. doi:10.1111/1745-9133.12044
- Beyond the Hype: AI in your SOC. (2020). <https://www.ibm.com/security/artificial-intelligence>
- Bindra, P., Kshirsagar, M., Ryan, C., Vaidya, G., Gupta, K. K., & Kshirsagar, V. (2021). *Insights into the Advancements of Artificial Intelligence and Machine Learning, the Present State of Art, and Future Prospects: Seven Decades of Digital Revolution*. 10.1007/978-981-16-0878-0_59
- Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). *YOLOv4: Optimal Speed and Accuracy of Object Detection*. ArXiv.
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. Advance online publication. doi:10.1023/A:1010933404324
- Chakraborty, P., Adu-Gyamfi, Y. O., Poddar, S., Ahsani, V., Sharma, A., & Sarkar, S. (2018). Traffic Congestion Detection from Camera Images using Deep Convolution Neural Networks. *Transportation Research Record: Journal of the Transportation Research Board*, 2672(45), 222–231. Advance online publication. doi:10.1177/0361198118777631
- Chaudhari, V. A., Kshirsagar, V., & Nagori, M. (2018, July). Integrating Sentiment Analysis and User Descriptors with Ratings in Sightseer Recommender System. *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. doi:10.1109/ICCCNT.2018.8494035
- Chidepatil, A., Bindra, P., Kulkarni, D., Qazi, M., Kshirsagar, M., & Sankaran, K. (2021). From Trash to Cash: How Blockchain and Multi-Sensor-Driven Artificial Intelligence Can Transform Circular Economy of Plastic Waste? *Administrative Sciences*, 10(2), 23. doi:10.3390/admsci10020023
- Chitode, K., & Nagori, M. (2013). A Comparative Study of Microarray Data Analysis for Cancer Classification. *International Journal of Computers and Applications*, 81(15), 14–18. Advance online publication. doi:10.5120/14198-2392
- Cloud AutoML Custom Machine Learning Models. (n.d.). Retrieved June 26, 2022, from <https://cloud.google.com/automl>
- Courtois, J.-P. (2019, August 7). *Harnessing the power of AI to transform agriculture*. Official Microsoft Blog. <https://blogs.microsoft.com/blog/2019/08/07/harnessing-the-power-of-ai-to-transform-agriculture/>
- Dickmanns, E. D., & Mysliwetz, B. D. (1992). Recursive 3-D road and relative ego-state recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2), 199–213. Advance online publication. doi:10.1109/34.121789
- Digital Agriculture: Farmers in India are using AI to increase crop yields. (2017). Retrieved July 24, 2021, from 40. <https://news.microsoft.com/en-in/features/ai-agriculture-icrisat-upl-india/>
- Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Ferrer, C. C. (2019). *The Deepfake Detection Challenge (DFDC) Preview Dataset*. 10.48550/ARXIV.1910.08854

- Elshaw, R., Al-Mallah, M. H., & Sakr, S. (2019). On the interpretability of machine learning-based model for predicting hypertension. *BMC Medical Informatics and Decision Making*, 19(1), 146. Advance online publication. doi:10.1186/s12911-019-0874-0 PMID:31357998
- Explainable AI (XAI) Market. (2021). <https://www.nextmsc.com/report/explainable-ai-market>
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5). Advance online publication. doi:10.1214/aos/1013203451
- Galkin, F., Mamoshina, P., Aliper, A., Putin, E., Moskalev, V., Gladyshev, V. N., & Zhavoronkov, A. (2020). Human Gut Microbiome Aging Clock Based on Taxonomic Profiling and Deep Learning. *iScience*, 23(6), 101199. Advance online publication. doi:10.1016/j.isci.2020.101199 PMID:32534441
- Garreau, D., & Luxburg, U. (2020). Explaining the Explainer: A First Theoretical Analysis of LIME. *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 1287–1296.
- Garson, G. D. (1991). Interpreting Neural Network Connection Weights. *Artif Intell Expert*, 6, 47–51.
- Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking Inside the Black-box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics*, 24(1), 44–65. Advance online publication. doi:10.1080/10618600.2014.907095
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 27). Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf>
- Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research*, 62, 729–754. doi:10.1613/jair.1.11222
- Green, D., & Kern, H. (2021). Modeling heterogeneous treatment effects in large-scale experiments using Bayesian Additive Regression Trees. *Public Opinion Quarterly*, 76(3), 491–511. doi:10.1093/poq/nfs036
- Groff, E., & LaVigne, N. (2002). Forecasting the Future of Predictive Crime Mapping. *Crime Prevention Studies*, 13, 26–57.
- Guardo, E., de Stefano, S., la Corte, A., Sapienza, M., & Scatà, M. (2018). A Fog Computing-based IoT Framework for Precision Agriculture. *Journal of Internet Technology*, 19(5), 1401–1411.
- Hamm, C. A., Wang, C. J., Savic, L. J., Ferrante, M., Schobert, I., Schlachter, T., Lin, M., Duncan, J. S., Weinreb, J. C., Chapiro, J., & Letzen, B. (2019). Deep learning for liver tumor diagnosis part I: Development of a convolutional neural network classifier for multi-phasic MRI. *European Radiology*, 29(7), 3338–3347. Advance online publication. doi:10.1007/s00330-019-06205-9 PMID:31016442
- Hansen, N. (2019, July 13). A global surrogate assisted CMA-ES. *Proceedings of the Genetic and Evolutionary Computation Conference*. doi:10.1145/3321707.3321842
- He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017, October). Mask R-CNN. *2017 IEEE International Conference on Computer Vision (ICCV)*. doi:10.1109/ICCV.2017.322
- High, R. (2012). *The Era of Cognitive Systems: An Inside Look at IBM Watson and How it Works*. Redguides for Business Leaders.
- How to Feed the World in 2050. (2009). https://www.fao.org/fileadmin/templates/wsfs/docs/expert_paper/How_to_Feed_the_World_in_2050.pdf
- Hubel, D., & Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1), 106–154. doi:10.1113/jphysiol.1962.sp006837 PMID:14449617
- Insider Threat Report. (2017). <https://crowdresearchpartners.com/portfolio/insider-threat-report/>
- Johns, K. (2020). *The top 12 Artificial Intelligence(AI) trends to watch out for in 2021*. ISHIR. Retrieved July 26, 2021, from <https://www.ishir.com/blog/9375/the-top-12-artificial-intelligenceai-trends-to-watch-out-for-in-2021.htm>

- Kapishnikov, A., Bolukbasi, T., Viegas, F., & Terry, M. (2019, October). XRAI: Better Attributions Through Regions. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. doi:10.1109/ICCV.2019.00505
- Kim, B., Seo, J., Jeon, S., Koo, J., Choe, J., & Jeon, T. (2019). *Why are Saliency Maps Noisy? Cause of and Solution to Noisy Saliency Maps*. ArXiv. doi:10.1109/ICCVW.2019.00510
- Kshirsagar, M., More, T., Lahoti, R., Adgaonkar, S., Jain, S., Ryan, C., & Kshirsagar, V. (2021). GREE-COCO: Green Artificial Intelligence Powered Cost Pricing Models for Congestion Control. *Proceedings of the 13th International Conference on Agents and Artificial Intelligence*. doi:10.5220/0010261209160923
- Kshirsagar, M., More, T., Lahoti, R., Adgaonkar, S., Jain, S., & Ryan, C. (2022). Rethinking Traffic Management with Congestion Pricing and Vehicular Routing for Sustainable and Clean Transport. *Proceedings of the 14th International Conference on Agents and Artificial Intelligence, 3(ICAART)*, 420–427. doi:10.5220/0010830300003116
- Kshirsagar, M., Vaidya, G., Rajguru, S., Jadhav, P., Kale, H., Shanmugam, N., Ryan, C., & Kshirsagar, V. (2022). DECART: Planning for Decarbonising Transport Sector with Predictive Analytics - An Irish Case Study. *Proceedings of the 11th International Conference on Smart Cities and Green ICT Systems*, 157–164. doi:10.5220/0011087100003203
- Kshirsagar, M., Vaidya, G., Yao, Y., Kasar, S., Ryan, C. (2022). IntelliMedChain: Knowledge Driven and Blockchain Powered Data Sharing Framework for Smart Healthcare. Authorea. 10.22541/au.165226321.19441546/v1
- Kshirsagar, V., & Joshi, M. (2015). Comparative Analysis of Various Classifiers for Performance Improvement in Intrusion Detection System by Reducing the False Positives. *International Journal of Computer Science and Information Technologies*, 5(5), 4825–4828.
- Kshirsagar, V., & Joshi, M. (2016). Rule Based Classifier Models For Intrusion Detection System. *International Journal of Computer Science and Information Technologies*, 7(1), 367–370.
- Lapuente-Santana, Ó., van Genderen, M., Hilbers, P. A. J., Finotello, F., & Eduati, F. (2021). *Interpretable systems biomarkers predict response to immune-checkpoint inhibitors*. Patterns. doi:10.1016/j.patter.2021.100293
- Lee, S., Lee, J., Lee, J., Park, C.-K., & Yoon, S. (2018). *Robust Tumor Localization with Pyramid Grad-CAM*. ArXiv.
- Liu, Z., Zheng, T., Xu, G., Yang, Z., Liu, H., & Cai, D. (2020). Training-Time-Friendly Network for Real-Time Object Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07), 11685–11692. Advance online publication. doi:10.1609/aaai.v34i07.6838
- López de Mántaras, R. (2018). The Future of AI: Toward Truly Intelligent Artificial Intelligences. In *Towards a New Enlightenment? A Transcendent Decade*. Academic Press.
- Lorente, M. P. S., Lopez, E. M., Florez, L. A., Espino, A. L., Martínez, J. A. I., & de Miguel, A. S. (2021). Explaining Deep Learning-Based Driver Models. *Applied Sciences (Basel, Switzerland)*, 11(8), 3321. Advance online publication. doi:10.3390/app11083321
- Machine Learning Market. (2020). <https://www.marketresearchfuture.com/reports/machine-learning-market-2494>
- Magenta. (n.d.). Retrieved June 26, 2022, from <https://magenta.tensorflow.org/>
- Mandal, V., Mussah, A. R., Jin, P., & Adu-Gyamfi, Y. (2020). Artificial Intelligence-Enabled Traffic Monitoring System. *Sustainability*, 12(21), 9177. Advance online publication. doi:10.3390/su12219177
- Marin, G. A. (2005). Network Security Basics. *IEEE Security and Privacy Magazine*, 3(6), 68–72. Advance online publication. doi:10.1109/MSP.2005.153
- Molnar, C. (2019). *Interpretable Machine Learning: A Guide for Making Black-box Models Explainable* (1st ed.). Lulu.
- Moravec, H. P. (1990). The Stanford Cart and the CMU Rover. In *Autonomous Robot Vehicles*. Springer. doi:10.1007/978-1-4613-8997-2_30

- Nagori, M., Jachak, R. S., & Chaudhari, P. P. (2019). A framework for segregating solid waste by employing the technique of image annotation. *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*. doi:10.1109/ICACCP.2019.8882932
- Nagori, M., Kulkarni, A., Keskar, K., Bhale, S., Sharangdhar, S., & Kshirsagar, V. (2018, April 29). Natural Language Processing for designing voice assistant for healthcare systems. *Sixth International Conference on Advances in Computing Communication and Information Technology CCIT 2018*. doi:10.15224/978-1-63248-149-8-19
- Nagori, M., Mutkule, S., & Sonarkar, P. (2013). Detection of Brain Tumor by Mining fMRI Images. *International Journal of Advanced Research in Computer and Communication Engineering*, 2(4), 1718–1722.
- Nagori, M., Patil, A., Deshmukh, S., Vaidya, G., Rahangdale, M., Kulkarni, C., & Kshirsagar, V. (2020). Mutichain Enabled EHR Management System and Predictive Analytics. *Smart Innovation. Systems and Technologies*, 165, 179–187. doi:10.1007/978-981-15-0077-0_19
- Niestadt, M., Debyser, A., Scordamaglia, D., & Pape, M. (2019). *Briefing*. European Parliamentary Research Service.
- Porto, R., Molina, J. M., Berlanga, A., & Patricio, M. A. (2021). Minimum Relevant Features to Obtain Explainable Systems for Predicting Cardiovascular Disease Using the Statlog Data Set. *Applied Sciences (Basel, Switzerland)*, 11(3), 1285. Advance online publication. doi:10.3390/app11031285
- Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. P., Shyu, M.-L., Chen, S.-C., & Iyengar, S. S. (2019). A Survey on Deep Learning. *ACM Computing Surveys*, 51(5), 1–36. Advance online publication. doi:10.1145/3234150
- Qazi, M., Kulkarni, D., & Nagori, M. (2020). Proof of Authenticity-Based Electronic Medical Records Storage on Blockchain. *Smart Innovation. Systems and Technologies*, 165, 297–306. doi:10.1007/978-981-15-0077-0_31
- Radford, A., Metz, L., & Chintala, S. (2015). *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*. arXiv. 10.48550/ARXIV.1511.06434
- Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6), 1137–1149. Advance online publication. doi:10.1109/TPAMI.2016.2577031 PMID:27295650
- Revenues from the Artificial Intelligence (AI) software Market worldwide from 2018 to 2025. (2020). <https://www.statista.com/statistics/607716/worldwide-artificial-intelligence-market-revenues>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August 13). “Why Should I Trust You?” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. doi:10.1145/2939672.2939778
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128(2), 336–359. Advance online publication. doi:10.1007/s11263-019-01228-7
- Shickel, B., Loftus, T. J., Adhikari, L., Ozrazgat-Baslanti, T., Bihorac, A., & Rashidi, P. (2019). DeepSOFA: A Continuous Acuity Score for Critically Ill Patients using Clinically Interpretable Deep Learning. *Scientific Reports*, 9(1), 1879. Advance online publication. doi:10.1038/s41598-019-38491-0 PMID:30755689
- Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*. ArXiv.
- Sundararajan, M., Taly, A., & Yan, Q. (2017). *Axiomatic Attribution for Deep Networks*. ArXiv.
- Threat Hunting Report. (2018). <https://crowdresearchpartners.com/portfolio/threat-hunting-report/>
- Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2018). Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. *Communications of the ACM*, 62(1), 96–104. doi:10.1145/3292039
- Tobin, S., Jayabalasingham, B., Huggett, S., & de Kleijn, M. (2020). A brief historical overview of artificial intelligence research. *Information Services & Use*, 39(4), 291–296. Advance online publication. doi:10.3233/ISU-190060

Tolido, R., Thieullent, A.-L., Linden, G. van der, Frank, A., Delabarre, L., Buvat, J., Theisler, J., Cherian, S., & Khemka, Y. (2019). *Reinventing Cybersecurity with Artificial Intelligence: The new frontier in digital security*. Academic Press.

Vaidya, G., Ilg, L., Kshirsagar, M., Naredo, E., & Ryan, C. (2022). HyperEstimator: Evolving Computationally Efficient CNN Models with Grammatical Evolution. *19th International Conference on Smart Business Technologies*, 57–68. doi:10.5220/0011324800003280

Weinreb, D., & Moon, D. (1981). The Lisp Machine manual. *ACM SIGART Bulletin*, 78(78), 10. Advance online publication. doi:10.1145/1056737.1056738

Weizenbaum, J. (1966). ELIZA—A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36–45. Advance online publication. doi:10.1145/365153.365168

Wu, H., Ruan, W., Wang, J., Zheng, D., Liu, B., Geng, Y., Chai, X., Chen, J., Li, K., Li, S., & Helal, S. (2021). Interpretable Machine Learning for COVID-19: An Empirical Study on Severity Prediction Task. *IEEE Transactions on Artificial Intelligence*. doi:10.1109/TAI.2021.3092698

Yang, C., Rangarajan, A., & Ranka, S. (2018). *Global Model Interpretation via Recursive Partitioning*. ArXiv. doi:10.1109/HPCC/SmartCity/DSS.2018.00256

Yao, Y., Kshirsagar, M., Vaidya, G., Ducrée, J., & Ryan, C. (2021). Convergence of Blockchain, Autonomous Agents, and Knowledge Graph to Share Electronic Health Records. *Frontiers in Blockchain*, 0, 13. doi:10.3389/fbloc.2021.661238

Završnik, A. (2020). Criminal justice, artificial intelligence systems, and human rights. *ERA Forum*, 20(4). doi:10.1007/s12027-020-00602-0

Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., Lyons, T., Manyika, J., Niebles, J. C., Sellitto, M., Shoham, Y., Clark, J., & Perrault, R. (2021). *The AI Index 2021 Annual Report*. Academic Press.

Zhao, G., Zhou, B., Wang, K., Jiang, R., & Xu, M. (2018). *Respond-CAM: Analyzing Deep Models for 3D Imaging Data by Visualizations*. 10.1007/978-3-030-00928-1_55

Meghana Kshirsagar is a Postdoctoral Researcher with Biocomputing and Developmental Systems Research Group and Lero, the Science Foundation Ireland Research Centre for Software since Sept. 2019 where she is currently co supervising on a PhD thesis. Her research interests include all domains of machine learning, artificial intelligence and blockchains with a particular focus on developing and deploying forecasting and prediction models for medical diagnostic systems, neural data science, knowledge graphs and visualization in data science. Previously, she was an Associate and Assistant Professor with Computer Science and Engineering Department at Government Engineering College Aurangabad, India.

Krishn Kumar Gupta received B.Sc. (Hons) degree in Electronics and M.Sc. in Informatics from the University of Delhi (DU), India in 2017 and 2019, respectively. He is currently pursuing a PhD degree in the domain of machine learning from the Department of Electrical and Electronic Engineering at the Technological University of the Shannon: Midlands Midwest (TUS), Ireland. He is currently working under the Science Foundation Ireland (SFI) funded project 'Automatic Design of Digital Circuit' (ADDC), based in the University of Limerick (UL) and Technological University of the Shannon, Ireland. He was a research student at the DU Innovation project, Delhi, and a RAWSC student at the Inter-University Centre for Astronomy and Astrophysics (IUCAA), Pune. He was awarded the prestigious summer research fellowship by the Indian Academy of Sciences (IASc), Bangalore, India in the year 2016 and 2018. His current research interest includes digital circuits design and testing, FPGA design, machine learning and evolutionary computation.

Gauri Vaidya is a PhD student at University of Limerick. She completed her Bachelors in Computer Science and Engineering Degree Program from Government College of Engineering Aurangabad, India in Aug 2020. She was a Trainee-Analyst, working as a salesforce developer with Principal Global Services, India for 11 months. Her research interests include blockchains, knowledge graphs, machine learning, artificial intelligence, and grammatical evolution.

Conor Ryan is a Professor of Machine Learning in the Computer Science and Information Systems (CSIS) department, a Funded Investigator within Lero (the Irish Software Research Centre) and a Science Foundation of Ireland Principal Investigator. He was a Fulbright Scholar at the Massachusetts Institute of Technology in 2013/14 and was CTO of NVMDurance, a company that optimized Flash Memory products, until it had a successful exit in early 2018. He is the inventor of Grammatical Evolution, one of the most widely used Automatic Programming systems.

Joe Sullivan received a BEng. in Electronic Engineering from the University of Brighton in 1994 and received his doctorate at the University of Limerick in 2013. Sullivan is CTO of NVMDurance, a company dedicated to extending the life of NAND memory. He became involved with Machine Learning as a solution to NV memory parameter optimisation in 2001 and has many published papers and 10 US patents in this field. Joe also lectures in Electronic Engineering at Limerick Institute of Technology and is a PI in the IDEAM Institutes and an academic member of the Lero Research Centre for Software at UL.

Vivek Kshirsagar is an Associate Professor in the Computer Science and Engineering Department at Government College of Engineering Aurangabad. He is Head of the Department for the last 13 years. His research areas are Cyber-Physical Systems, Computer Networking, and Machine Learning.