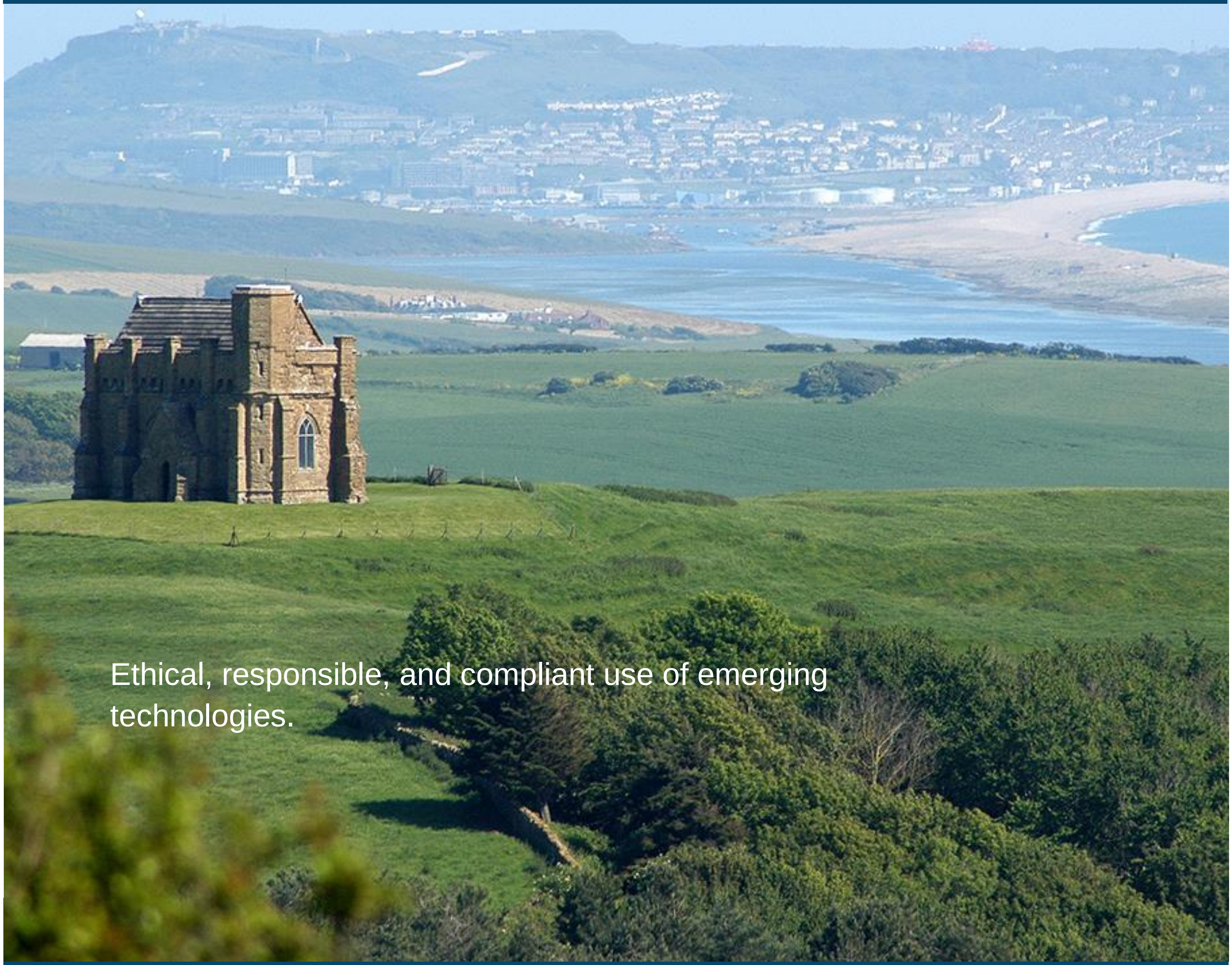
Policy for the use of AI, Automation & Algorithmic Data Processing in Dorset Council



Ethical, responsible, and compliant use of emerging technologies.

Contents

Dorset
Council

Content

1. Summary

This policy provides guidelines for the responsible use of Artificial Intelligence, Intelligent Automation, or algorithmic data processing in Dorset Council.

The policy aims to ensure that the use of these technologies is;

- **Ethical – socially & morally doing the right thing**
- **Responsible – being careful about what we use it for**
- **Compliant – following the relevant standards and rules**
- **Transparent – about what and how we use it**
- **and never compromise the security or privacy of our customers or staff**

The policy applies to all users of IT systems to process data in a way that has a direct or indirect impact on or is accessed by the council's customers or staff.

The policy outlines the expectations when using these technologies, as well as the potential risks and challenges they pose.

The policy also provides guidance and considerations for people who intend to use these technologies, such as seeking approval, reviewing accuracy, disclosing content, integrating with other tools, and providing citations.

2. Overview

The use of AI, Automation, and other emerging technologies present significant opportunities to enhance both efficiency and customer experience. These technologies will enable the Authority to streamline processes, reduce manual tasks, and improve accuracy, leading to increased productivity and ultimately cost savings. Moreover, these technologies can greatly enhance customer experience by providing personalised and timely services. AI can analyse customer data to predict needs and preferences, allowing for more tailored interactions. Automation can ensure that customer inquiries are handled promptly and accurately, improving overall satisfaction.

Artificial Intelligence (AI), Intelligent Automation and Algorithmic Data Processing in the context of this policy are all terms for systems that through the processing of data have a direct or indirect impact on or are directly accessed by customers and staff of Dorset Council (DC). This policy is not limited to any specific system and could include the use of ‘line of business’ applications with integrated components used to process the data in our custody to elicit an outcome, automate a response or influence a decision that could have a material impact on the data subject/s. For ease of reference, AI will be used as a general term for systems that process data in this way.

A subset of deep learning, some of these systems can produce new content based on user inputs or prompts, these models can generate language, code, and images. They are connected to huge data sets that have been ‘trained’ or connected in a neural network and can predict or produce outputs that are similar to the natural language inputs, but not identical. In most free to use services any data you have entered into the model is consumed and further trains the model itself.

This technology is advancing rapidly, and new applications are being developed every day. However, the pace at which they are evolving, and their adoption presents certain risks and challenges. These models can inadvertently reveal sensitive information from the training data, perpetuate and amplify biases present in the data, be targeted by malicious actors who seek to manipulate the outputs or occasionally invent content to fill gaps in its knowledge (hallucinations). It is important to carefully consider the ethics, bias, and source of data in its responses, as well as the sensitivity of the data being entered.

3. Context

Data forms the foundation of our organisation and facilitates our ability to deliver high quality, efficient, customer focused services at the place and time they’re needed. We hold an enormous amount of data that tells us where we were, where we are, and where we’re going and when put together correctly, provides powerful forecasting tools and actionable insights. This is what we mean by ‘data led organisation’ and can be explored in more detail in the Dorset Council Data Strategy.

The enormous growth of our digital data assets has in turn led to the need to invest in a variety of systems to manage them. These systems have historically been isolated silo's providing very little beyond the core line of business capability and only removing redundant data at the point of decommissioning. New systems that process data within or on behalf of Dorset Council and its clients present some key risks that will need to be addressed prior to procuring, developing, deploying, integrating and/or administering them. These risks include data breaches, bias in base data, attacks by malicious actors, and use of redundant or outdated data.

We all have a responsibility to protect the sensitive information we have access to, and to ensure that it is used in a transparent, fair, and lawful manner. The use of these technologies must be carefully managed to ensure that it does not compromise the security or privacy of this information.

At the same time, they have the potential to drive innovation and improve public services. They can be used to automate routine tasks, generate insights from data, and provide personalised services to citizens. By adopting a Pro-Innovation approach to regulation, Local Government organisations can strike a balance between opportunity and risk.

4. Purpose

The purpose of this policy document is to enable us to take advantage of emerging technologies in a safe, secure and responsible way. It will also outline the potential risks and impact of inappropriate use and regulations that must be considered by Councillors, council employees, contractors, developers, vendors, temporary staff, consultants or other third parties, hereinafter referred to as 'users'.

This policy is designed to ensure that our use of AI is ethical, complies with all applicable laws, regulations, and council policies, and complements the council's existing acceptable use, information and security policies.

5. Use

This policy applies to all users with access to IT systems, whether through council-owned devices or BYOD (bring your own device) in pursuit of council activities.

Use of AI systems needs to be in a manner that promotes fairness and avoids bias to prevent discrimination and promote equal treatment and be in such a way as to contribute positively to the council's goals and values.

Users may access AI tools for work-related purposes subject to adherence to the following policy. This includes tasks such as **generating** text or content for reports, emails, presentations, images, and customer communications.

Where the user intends to **enter** data into an AI interface, great care must be taken to ensure beyond doubt that the source data contains no sensitive or Personally Identifiable Information (PII), whether to summarise, generate content or influence outcomes and/or decisions without first following the approved process of Change Proposal, Assessment and Solution Recommendation by appropriate Subject Matter Experts (SME) as assigned by the Design Advisor & Alignment Group.

Guidance and Learning Pathways are being developed and will, at different levels and contexts be mandated to ensure use of AI is appropriate.

Where algorithmic processing of any kind is carried out which directly or indirectly impacts or is accessed by a customer, whether considered AI or not, must be transparent and explainable. The Data Protection Officer will be able to guide you through the Algorithmic Transparency Recording Standard (ATRS) which consists of an assessment carried out by both the technology provider and the customer on behalf of the Cabinet Office and Department for Science Innovation & Technology. On completion of the assessment an entry will be made on a publicly accessible register to provide transparency of our processing.

When engaging with any AI particular attention should be given to Governance, Vendor Practices, Copyright, Accuracy, Confidentiality, Ethical Use, Environmental Impact, Disclosure, and Integration with other tools.

5.1 Governance

Before using AI technology for any purpose where Council or Customer data is consumed, processed, or stored and where the output could have material impact or potential risk of data breach, users must first seek approval from the council's Operational Information Governance Group (OIGG). The intended use, the reason for use, and the expected information to be input as well as the generated output and its distribution will need to be provided. This will form part of the process described above and be undertaken with the support of the appropriate SMEs. Proposals will be signed off by the appropriate Information Asset Owner, incorporating any requirements, recommendations or comments raised by the OIGG.

5.2 Vendors & Procurement

Any use of AI technology in pursuit of council activities should be done with full acknowledgement of the policies, practices, terms & conditions of developer/vendor. Before engaging in any procurement activity where it is known or intended to incorporate AI technology, procurement processes will incorporate a supported AI Vendor Acceptance Checklist. This will be made available, and support provided by the Automation & AI Team during a procurement/commissioning process.

5.3 Copyright

Users must adhere to copyright laws when utilising AI. It is prohibited to use AI to generate content that infringes upon the intellectual property rights of others, including but not limited to copyrighted material. **If a user is unsure whether a particular use of AI constitutes copyright infringement, they should contact the legal advisor or Information Governance Team before using AI.**

5.4 Accuracy

All information generated by AI must be reviewed and edited for accuracy prior to use. Users of AI are responsible for reviewing output and are accountable for ensuring the accuracy of generated output before use/release. Use of AI in sensitive, regulated, and legal processes where the output cannot be guaranteed or monitored must consider the impact of the output. It is important that officers provide accurate information. Officers should take particular care in circumstances such as providing advice to a vulnerable client. However, they should also be mindful about the importance of providing accurate information to residents when discharging any council duty, such as provision of information about services such as bin collections or missed bins. **If a user has any doubt about the accuracy of information generated by AI or the appropriate context, they should not use it and seek advice from their line manager who should then consult the Automation& AI team if they are not sure of the appropriate course of action.**

5.5 Confidentiality

Confidential and personal information must not be entered into an AI tool, unless the user is certain it will not enter the public domain or be accessible to unauthorised internal users. Users must complete the Data Protection Screening Tool, full Impact Assessment (if required) and follow all applicable data privacy laws and organisational policies when using AI. **If a user has any doubt about the confidentiality of information, they should not use AI to process it.**

5.6 Ethical Use

AI must be used ethically and in compliance with all applicable legislation, regulations, and organisational policies. Users must not use AI to generate content that is discriminatory, offensive, or inappropriate. Where it is not possible or intended to have a Human in the Loop (HITL) the appropriate level of testing and audit regime should be applied.

The application of AI tools can and will have an impact on ways of working. An appropriate Equality Impact Assessment (EqIA) should be carried out before commissioning any use case. **If there are any doubts about the appropriateness of using AI in a particular situation, users should consult with their manager or Information Governance Team.**

5.7 Environmental Sustainability

AI must be used sustainably, and users should consider the environmental impact of their use. The environmental impacts of AI are realised through its development (training phase), operation (inference phase) and in its applications. Negative development and operational impacts include high energy consumption and water use, a significant carbon footprint, and the creation of electronic waste – and these may grow as more sophisticated models are developed and usage grows. In application it can be either environmentally positive or negative, but it is possible that its applications can support sustainability through optimising production processes and automating tasks in ways which enables more efficient, less carbon-intensive processes or a shift to low carbon technologies.

Growing understanding of ‘green AI’ (contrasted with ‘red AI’) aims to reduce reliance on ever-larger models, data, and compute to limit AI’s negative environmental impacts – but there are as yet no clear standards or certifications. Nevertheless, there are a clear set of factors that should be considered when using or procuring AI-powered services, including:

- Sensible use: Avoiding use of AI services where less compute-intensive (and thereby carbon-intensive) methods would suffice
- Carbon transparency: Seek clarity about the specific impacts of particular AI services using existing tracker or impact dashboard tools
- Mitigation: Prefer use or procurement of AI-services which are trained and operated through data centres powered by clean energy; which use of low-carbon cooling methods and heat recovery technologies; which optimise the energy efficiency of data centre management, hardware, and the algorithmic/computational efficiency of AI model architecture, training and development; and which considers its whole-lifecycle impacts, from mineral extraction through to end-of-life treatment of electronic waste.

Use and application of AI services should carefully consider the environmental impacts of each of the development and operation of the AI services as well as the consequences of its specific application, in line with the council’s sustainability strategy and through standard use of our decision tool. Continuous monitoring and the pursuit of some mitigations will form part of broader action to mitigate the impacts of cloud computing and IT-related environmental impacts, but all users or procurers of specific applications should be mindful of their own role. This policy will be revised to reflect future relevant regulation, policy, or standards as such emerges.

5.8 Disclosure

Content produced via AI must be identified and disclosed as containing AI-generated information.

Example: ***Note:*** *This document contains content generated by Artificial Intelligence (AI). AI generated content has been reviewed by the author for accuracy and edited/revised where necessary. The author takes responsibility for this content.*

5.9 Integration

API and plugin tools enable access to AI and extended functionality for other services to improve automation and productivity outputs. **Users intending to develop Automations and Integrations that call any AI service should also complete the AI Use Case Assessment toolkit to ensure appropriate consideration and mitigation of risk. This should include but is not limited to:**

- Adversarial testing
- Human in the Loop (HITL)
- Prompt engineering
- "Know Your Customer" (KYC) – Engage early and often
- Constrain/control user inputs and limit output tokens
- Allow users to report issues
- Understand and communicate limitations

API and plugin tools must be rigorously tested for:

- Moderation – to ensure the model properly handles hate, discriminatory, threatening, etc. inputs appropriately.
- Factual responses – provide a ground of truth for the API and review responses.

6. Risks

Use of many emerging technologies such as GenAI, Automation and systems used to process data carry inherent risks. A comprehensive risk assessment should be conducted for any project or process where use of these products is proposed. The risk assessment should consider potential impacts including legal compliance; bias and discrimination; security (including technical protections and security certifications); data sovereignty and protection; knowledge about source data, and accountability.

6.1 Legal Compliance

Data entered into free-to-access AI may enter the public domain. This can release non-public information and breach regulatory requirements, customer, or vendor contracts, or compromise intellectual property. Any release of private/personal information without the authorisation of the information's owner could result in a breach of relevant data protection laws. Use of AI to compile content may also infringe on regulations for the protection of intellectual property rights. **Users should ensure that their use of any AI complies with all applicable laws and regulations and with council policies.**

6.2 Bias & Discrimination

AI may make use of and generate biased, discriminatory, or offensive content. **Users should use AI responsibly and ethically, in compliance with council policies and applicable laws and regulations.**

6.3 Security

AI may store sensitive data and information, which could be at risk of being breached or hacked. The council must assess technical protections and security certification of AI before use. **If a user has any doubt about the security of information input into AI, they should not use it.**

6.4 Data Protection & Sovereignty

While an AI platform may be hosted internationally, under data sovereignty rules information created or collected in the originating country will remain under jurisdiction of that country's laws. The reverse also applies. If information is sourced from AI hosted overseas, the laws of the source country regarding its use and access may apply. **AI service providers should be assessed for data sovereignty practice by any organisation wishing to use their products.**

6.5 Knowledge about Source Data

AI solutions are more likely to generate unreliable outputs or 'confident inaccuracy' if the source data contains outdated information. There is an increased risk of data breaches if the source data has overly generous permissions. **Users must be fully aware of the contents of the source data, especially when its quality is difficult to assess because it is held in an uncontrolled file/folder structure.**

6.6 Accountability

AI generated content will add more sensitive data that needs to be managed. All assets must be updated on the council's Information Asset Register to account for new AI content. Before an AI solution is procured or deployed, the Information Asset Owner who will be responsible for the asset must make any required amendments to the relevant privacy notice and decide whether a Data Protection Impact Assessment should be carried out, in consultation with the Data Protection Officer. The future management of generated content must be considered upfront and built into the solution's design.

7. Enforcement

In line with the ICT Acceptable Use Policy if there is evidence to suggest that users have wilfully or negligently failed to abide by these policy requirements, Dorset Council has the right to investigate which may result in disciplinary action in line with the Council's disciplinary procedure.

If there is reason to believe that a data breach has occurred, the Information Compliance Team must be advised urgently, in line with the Council's Data Breach Policy.

8. Review

This policy will be reviewed periodically and updated as necessary to ensure continued compliance with all applicable legislation, regulations, and organisational policies, but at a minimum every three years.

9. Acknowledgments

By using Automation, GenerativeAI, or any products that process client or council data the user acknowledges that they have read and understood these guidelines, including the risks associated with their use and have taken appropriate mitigating actions.

This policy has been based on guidance previously prepared by ALGIM (Aotearoa - New Zealand) and SOCITM (UK).

ALGIM

Socitm

Policy Author: Brian Hole, Digital Programme Manager

Policy owner: Marc Eyre, Service Manager for Assurance (Chair of Operational Information Governance Group)

Date approved: Strategic Information Governance Board – 19 September 2024

Review date: September 2027

Glossary

Adversarial Testing

Adversarial testing is a method used to evaluate machine learning models by intentionally exposing them to malicious or harmful inputs to identify vulnerabilities and weaknesses. This process helps in understanding how models behave under attack and guides improvements to make them more robust and secure.

Algorithm

A set of instructions used to perform tasks (such as calculations and data analysis) usually using a computer or another smart device.

Algorithmic Bias

AI systems can have bias embedded in them, which can manifest through various pathways including biased training datasets or biased decisions made by humans in the design of algorithms.

Artificial Intelligence (AI)

The UK Government's 2023 policy paper on 'A pro-innovation approach to AI regulation' defined AI, AI systems or AI technologies as "products and services that are 'adaptable' and 'autonomous'." The adaptability of AI refers to AI systems, after being trained, often developing the ability to perform new ways of finding patterns and connections in data that are not directly envisioned by their human programmers. The autonomy of AI refers to some AI systems that can make decisions without the intent or ongoing control of a human.

Artificial General Intelligence (AGI)

Sometimes known as general AI, strong AI or broad AI, this often refers to a theoretical form of AI that can achieve human-level or higher performance across most cognitive tasks. See also Superintelligence.

Artificial Neural Network

A computer structure inspired by the biological brain, consisting of a large set of interconnected computational units ('neurons') that are connected in layers. Data passes between these units as between neurons in a brain. Outputs of a previous layer are used as inputs for the next, and there can be hundreds of layers of units. An artificial neural network with more than 3 layers is considered a deep learning algorithm. Examples of artificial neural networks include Transformers or Generative adversarial networks.

Automated Decision-Making

A term that the Office for AI, within the Department for Science, Innovation and Technology, refers to in an Ethics, Transparency and Accountability Framework for Automated decision-making as “both solely automated decisions (no human judgement involved) and automated assisted decision-making (assisting human judgement).” AI systems are increasingly being used by the public and private sector for automated decision-making.

Computer Vision

This focuses on programming computer systems to interpret and understand images, videos and other visual inputs and take actions or make recommendations based on that information. Applications include object recognition, facial recognition, medical imaging analysis, navigation and video surveillance.

Deep Learning

A subset of machine learning that uses artificial neural networks to recognise patterns in data and provide a suitable output, for example, a prediction. Deep learning is suitable for complex learning tasks and has improved AI capabilities in tasks such as voice and image recognition, object detection and autonomous driving.

Deepfakes

Pictures and video that are deliberately altered to generate misinformation and disinformation. Advances in generative AI have lowered the barrier for the production of deepfakes.

Disinformation

Disinformation is the “deliberate creation and spreading of false and/or manipulated information that is intended to deceive and mislead people, either for the purposes of causing harm, or for political, personal or financial gain”. Advances in generative AI have lowered the barrier for the production of disinformation, misinformation, and deepfakes.

Fine-Tuning

Fine-tuning a model involves developers training it further on a specific set of data to improve its performance for a specific application.

Foundation Models

A machine learning model trained on a vast amount of data so that it can easily be adapted for a wide range of general tasks, including being able to generate outputs (generative AI). See also large language models.

Frontier AI

Defined by the Government Office for Science as 'highly capable general-purpose AI models that can perform a wide variety of tasks and match or exceed the capabilities present in today's most advanced models'. Currently, this primarily encompasses a few large language models including

- ChatGPT (OpenAI)
- Claude (Anthropic)
- and Bard (Google)

Generative AI

An AI model that generates text, images, audio, video or other media in response to user prompts. It uses machine learning techniques to create new data that has similar characteristics to the data it was trained on. Generative AI applications include chatbots, photo and video filters, and virtual assistants.

General-Purpose AI

Often refers to AI models that can be adapted to a wide range of applications (such as Foundation Models). See also narrow AI.

Hallucinations

Large language models, such as ChatGPT, are unable to identify if the phrases they generate make sense or are accurate. This can sometimes lead to inaccurate results, also known as 'hallucination' effects, where large language models generate plausible sounding but inaccurate text. Hallucinations can also result from biases in training datasets or the model's lack of access to up-to-date information.

Human in the Loop (HITL)

Human-in-the-loop (HITL) is a process where human judgment and feedback are integrated into automated systems, such as AI and machine learning models, to enhance their accuracy and relevance. By involving humans in the decision-making loop, HITL ensures that the systems can benefit from human expertise and intuition, leading to more reliable and realistic outcomes. This approach is widely used in various applications, including simulations, autonomous systems, and training scenarios, where human oversight is crucial for optimal performance.

Interpretability

Some machine learning models, particularly those trained with deep learning, are so complex that it may be difficult or impossible to know how the model produced the output. Interpretability often describes the ability to present or explain a machine learning system's decision-making process in terms that can be understood by humans. Interpretability is sometimes referred to as transparency or explainability.

Large Language Models (LLM)

A type of foundation model that is trained on vast amounts of text to carry out natural language processing tasks. During training phases, large language models learn parameters from factors such as the model size and training datasets. Parameters are then used by large language models to infer new content. Whilst there is no universally agreed figure for how large training datasets need to be, the biggest large language models (frontier AI) have been trained on billions or even trillions of bits of data. For example, the large language model underpinning ChatGPT 3.5 (released to the public in November 2022) was trained using 300 billion words obtained from internet text. See also natural language processing and foundation models.

Machine Learning

A type of AI that allows a system to learn and improve from examples without all its instructions being explicitly programmed (PN 633). Machine learning systems learn by finding patterns in training datasets. They then create a model (with algorithms) encompassing their findings. This model is then typically applied to new data to make predictions or provide other useful outputs, such as translating text. Training machine learning systems for specific applications can involve different forms of learning, such as supervised, unsupervised, semi-supervised and reinforcement learning.

Misinformation

The UK Government defines misinformation as "the inadvertent spread of false information". Advances in generative AI have lowered the barrier for the production of disinformation, misinformation, and deepfakes.

Narrow AI

Sometimes known as weak AI, these AI models are designed to perform a specific task (such as speech recognition) and cannot be adapted to other tasks. See also general-purpose AI.

Natural Language Processing (NLP)

This focuses on programming computer systems to understand and generate human speech and text. Algorithms look for linguistic patterns in how sentences and paragraphs are constructed and how words, context and structure work together to create meaning. Applications include speech-to-text converters, online tools that summarise text, chatbots, speech recognition and translations.

Open-Source

Open-source often means the underlying code used to run AI models is freely available for testing, scrutiny and improvement.

Prompt Engineering

Prompt engineering is the process of designing and refining prompts to effectively communicate with AI models, ensuring they generate the desired responses. This involves crafting specific and clear instructions, providing context, and sometimes iterating on the prompts to improve the quality of the output. By understanding the model's behaviour and capabilities, prompt engineers can optimize the interaction between humans and AI, making the AI's responses more accurate and useful

Reinforcement Learning

A way of training machine learning systems for a specific application. An AI system is trained by being rewarded for following certain 'correct' strategies and punished if it follows the 'wrong' strategies. After completing a task, the AI system receives feedback, which can sometimes be given by humans (known as 'reinforcement learning from human feedback'). In the feedback, positive values are assigned to 'correct' strategies to encourage the AI system to use them, and negative values are assigned to 'wrong' strategies to discourage them, with the classification of 'correct' and 'wrong' depending on a pre-established outcome. This type of learning is useful for tweaking an AI model to follow certain 'correct' behaviours, such as fine-tuning a chatbot to output a preferred style, tone or format of language.

Responsible AI

Often refers to the practice of designing, developing, and deploying AI with certain values, such as being trustworthy, ethical, transparent, explainable, fair, robust and upholding privacy rights.

Semi-Supervised Learning

A way of training machine learning systems for a specific application. An AI system uses a mix of supervised and unsupervised learning and labelled and unlabelled data. This type of learning is useful when it is difficult to extract relevant features from data and when there are high volumes of complex data, such as identifying abnormalities in medical images, like potential tumours or other markers of diseases. See also supervised learning, unsupervised learning, reinforcement learning and training datasets.

Supervised Learning

A way of training machine learning systems for a specific application. In a training phase, an AI system is fed labelled data. The system trains from the input data, and the resulting model is then tested to see if it can correctly apply labels to new unlabelled data (such as if it can correctly label unlabelled pictures of cats and dogs accordingly). This type of learning is useful when it is clear what is being searched for, such as identifying spam mail.

Training Datasets

The set of data used to train an AI system. Training datasets can be labelled (for example, pictures of cats and dogs labelled ‘cat’ or ‘dog’ accordingly) or unlabelled.

Transformers

Transformers have greatly improved natural language processing, computer vision and robotic capabilities and the ability of AI models to generate text. A transformer can read vast amounts of text, spot patterns in how words and phrases relate to each other, and then make predictions about what word should come next. This ability to spot patterns in how words and phrases relate to each other is a key innovation, which has allowed AI models using transformer architectures to achieve a greater level of comprehension than previously possible.

Unsupervised learning

A way of training machine learning systems for a specific application. An AI system is fed large amounts of unlabelled data, in which it starts to recognise patterns of its own accord. This type of learning is useful when it is not clear what patterns are hidden in data, such as in online shopping basket recommendations (“customers who bought this item also bought the following items”). See also semi-supervised learning, supervised learning and reinforcement learning and training datasets.