Northern Ireland Assembly

Research and Information Service
Research Paper

25/09/2023

Dr. Aaron McMurray

Artificial Intelligence, Machine Learning, Deep Learning and Large Language Models

NIAR 96-2024

This paper outlines the main differences between artificial intelligence, machine learning, deep learning and large language models. The regulatory approach of the United Kingdom and the potential challenges are also detailed. In addition, specific use cases for large language models in research are explored.

Paper 12/24 25 September 2023

Research and Information Service briefings are compiled for the benefit of MLAs and their support staff. Authors are available to discuss the contents of these papers with Members and their staff but cannot advise members of the general public. We do, however, welcome written evidence that relates to our papers and this should be sent to the Research and Information Service, Northern Ireland Assembly, Room 139, Parliament Buildings, Belfast BT4 3XX or e-mailed to Raisecsu@niassembly.gov.uk

Key Points

Machine learning and deep learning are fields within the broader field of Artificial Intelligence (AI) with deep learning being a sub-field of machine learning.

Machine learning is focused on the use of data and algorithms to imitate the ways in which humans typically learn, gradually improving accuracy over time.

Deep learning is based on the use of specific algorithmic structures called “artificial neural networks,” which replicate the structure of brain.

Large language models are models built from artificial neural networks which are able to process large amounts of text data.

ChatGPT (Chat Generative Pre-Trained Transformer) is a large language model based chatbot developed by OpenAI and launched on November 30 2022.

ChatGPT is built upon GPT-3.5 and GPT-4 which are part of OpenAI’s proprietary series of GPT models fine-tuned for conversational applications.

OpenAI states the GPT-4 hallucinates facts and makes reasoning errors; does not learn from experience; and is overly gullible in accepting obviously false statements from a user. Some of these issues present avenues to exploit GPT.

The House of Lords Artificial Intelligence Committee raised concerns around poorly representative data sets used to train AI systems; the monopolisation of data; and the consolidation of power by overseas technology companies.

A recent government white paper outlines a preference for a light-touch approach to regulation focused on “regulating the use, not the technology,” which contrasts the European Union’s movement towards a risk-based, legislative approach to AI.

AI tools pose challenges but also present opportunities to transform how we work. GPT models can make useful research assistants capable of answering questions relating to dense texts.

There will be challenges ahead in finding a balance between adopting a new technology that is clearly here to stay and facing up to the difficult questions that technology raises around privacy and digital rights.

Executive Summary

The UK’s National AI Strategy is committed to ensuring that “the UK remains an AI and science superpower” into the early 2030s and the UK Science and Technology Framework identified AI as one of 5 critical technologies of tomorrow.

The terms “machine learning,” and “deep learning” are commonly used in the context of AI. These terms are often used interchangeably however there are differences between the two. Both of these are fields within the broader field of AI with deep learning being a sub-field of machine learning.

Machine learning is focused on the use of data and algorithms to imitate the ways in which humans typically learn, gradually improving accuracy over time. Through the use of statistical methods, algorithms are “trained” to make classifications or predictions about various data sets.

Deep learning differs from machine learning in that it is based on the use of specific algorithmic structures called “artificial neural networks.” Artificial neural networks are a branch of machine learning models built on the principle of replicating the structure of brain.

Large language models are able to process large amounts of text data, typically scraped from the Internet. Large language models are built from artificial neural networks.

ChatGPT (Chat Generative Pre-Trained Transformer) is a large language model based chatbot developed by OpenAI and launched on 30 November 2022.

ChatGPT is built upon GPT-3.5 and GPT-4 which are part of OpenAI’s proprietary series of GPT models that are fine-tuned for conversational applications using supervised and reinforcement learning techniques. In the first week since its launch it had over 1 million users.

In describing its limitations OpenAI describes GPT-4 as “not fully reliable,” stating that it hallucinates facts and makes reasoning errors; lacks knowledge of events occurring after September 2021; does not learn from experience; and is overly gullible in accepting obviously false statements from a user. Some of these issues present avenues to exploit GPT.

There are also privacy and GDPR concerns surrounding the use of AI. The House of Lords Artificial Intelligence Committee raised concerns, noting that many of the data sets used to train AI systems were poorly representative of the wider population. The Committee also raised concerns relating to the monopolisation of data; the consolidation of power and influence by a select few overseas technology companies in the UK; and the risk posed to the home-grown AI start-up sector.

A recent government white paper appears to outline a preference for a light-touch approach to regulation focused on “regulating the use, not the technology,” which stands in stark contrast with the European Union’s apparent movement towards a risk-based approach legislative approach to AI that establishes obligations for providers and users depending on the level of risk the AI generates.

While AI tools pose challenges they also present opportunities to transform how we work. OpenAI’s GPT models can for instance can make useful research assistants capable of answering questions relating to dense texts.

There will be challenges ahead in finding a balance between adopting a new technology that is clearly here to stay and facing up to difficult questions that technology raises around privacy and digital rights.

Contents

Key Points 1

Executive Summary 3

1 Artificial Intelligence, Machine Learning and Deep Learning – What’s the Difference? 6

1.1 Artificial Intelligence 6

1.2 Machine Learning 8

1.3 Deep Learning and Neural Networks 10

2 Large Language Models 13

2.1 The Ghost in the Machine 16

2.2 ChatGPT’s Alter Ego “DAN” 18

2.3 Privacy Concerns and GDPR Challenges 20

3 A Regulatory Black Hole? 23

4 AI – What is it Good For? 28

5 AI as a Research Assistant 29

6 Conclusions 31

1 Artificial Intelligence, Machine Learning and Deep Learning – What's the Difference?

1.1 Artificial Intelligence

The UK's National AI Strategy is committed to ensuring that the UK "remains an AI and science superpower" into the early 2030s[1] and the UK Science and Technology Framework has identified AI as one of five critical technologies of tomorrow, essential to the government's strategic vision. AI is taking off in Northern Ireland too. PWC have predicted that Northern Ireland's GDP could see a boost equal to an approximate spending power increase of £2,000 a year by 2030 as a result of the impact of AI across the UK's commercial activities[2].

A number of firms in Northern Ireland are using AI and automation technologies and there is substantial academic AI capability in both the Ulster University and Queen's University with over 150 permanent staff involved in AI research across the two universities[3]. Ulster University has the largest computer science grouping on the island of Ireland and the 14th largest in the UK with AI being a main focus of the Ulster University's Computer Science research for over two decades[4]. Twelve percent of Ulster University's AI publications were in the top 10% citation percentiles in AI[5].

Within Queen's University research is conduced across a range of Faculties and Schools with several main centers of activity including the School of Mathematics and Physics, the School of Electronics Electrical Engineering and Computer Science and the Stratified Cancer Medicine Cluster. Almost ten

---

[1] Department for Digital Culture Media and Sport, 2021, National AI Strategy, Command Paper Number: 525, accessed: 11/08/2023.

[2] PwC, The economic impact of artificial intelligence on the UK economy, published: 06/2017, accessed: 18/09/2023.

[3] Invest NI, Northern Ireland, A location of choice for AI and Data Analytics, accessed: 30/08/2023.

[4] Matrix, (2019), Artificial Intelligence Research in Northern Ireland, accessed: 30/08/2023.

[5] Matrix, (2019), Artificial Intelligence Research in Northern Ireland, accessed: 30/08/2023.

percent of Queen's University Belfast AI publications were in the top 10% citation percentiles in AI[6].

A number of spin-outs have come from the Ulster University including[7]: Intelesens which uses AI in wireless wearable ECG and respiration detection systems; Heartsine Technologies which developed Automated External Defibrillators using defibrillation algorithms; axial3d which is a medical technology firm focused on global adoption of 3D printing in healthcare; and Datactics which develops software platforms to cleanse and reformat data for banking, finance, government and healthcare analytics.

Larger companies such as Allstate, Kainos, Citi, BT, PWC as well as Etain, RepKnight, Analytics Engines, Serafim and Big Motive are also involved in AI development. These established businesses are typically interested in using AI to improve their processes[8].

In what seems like a remarkably short time span AI has been woven into the fabric of our lives. It permeates diverse sectors from healthcare and finance to transportation and entertainment. Its influence is reshaping the landscape of technology and human interaction. AI will compel us to reconsider how prepared we are socially, culturally and legally for a world filled with AIs processing data and making decisions on our behalf.

Numerous definitions have been proposed for AI with the first real discussion around machine intelligence being proposed by Alan Turing who asked the question: “can machines think?”[9] In his paper, Computing Machinery and Intelligence, he discussed the conditions for considering a machine to be intelligent and argued that if it could successfully pretend to be human and

---

[6] Matrix, (2019), Artificial Intelligence Research in Northern Ireland, accessed: 30/08/2023.

[7] Matrix, (2019), Artificial Intelligence Research in Northern Ireland, accessed: 30/08/2023.

[8] Matrix, (2019), Artificial Intelligence Research in Northern Ireland, accessed: 30/08/2023.

[9] Alan Turing, 1950, Computing Machinery and Intelligence, Mind 49: 433-460.

convince a knowledgeable observer then it ought to be considered intelligent. This test is the now famous “Turing Test.”

The Turing test has been discussed at length since it was first introduced. It has been discussed by both researchers and lay persons[10] and the test has even made its way into the works of science fiction authors. The famous film, Blade Runner, based on the novel “Do Androids Dream of Electric Sheep?” features a world in which humans and robots are indistinguishable except through a series of questions invoking emotional responses. While the bar for intelligence set by the Turing test and modern sci-fi is focused on mimicry of human intelligence other definitions of AI are less concerned with human thinking and more concerned with the ability of AI to simply act with rationality. The House of Commons Science and Technology Committee, in a report on Algorithms in decision making, noted that while there is no single agreed definition of AI there are commonalities between definitions. The Committee drew from a range of sources to define AI broadly as a set of statistical tools and algorithms that combine to form, in part, intelligent software, enabling computers to simulate elements of human behaviour such as learning, reasoning and classification[11].

1.2 Machine Learning

The terms “machine learning,” and “deep learning” are commonly used in the context of AI. These terms are often used interchangeably however there are differences between the two. Both of these are fields within the broader field of AI with deep learning being a sub-field of machine learning.

Machine learning is focused on the use of data and algorithms to imitate the ways in which humans typically learn, gradually improving accuracy over time. Through the use of statistical methods, algorithms are “trained” to make classifications or predictions about various data sets. The term itself was coined

---

[10] John McCarthy, 2007, What is Artificial Intelligence, accessed: 07/08/2023.

[11] House of Commons Science and Technology Committee, 2018, Algorithms in decision making, published: 15/05/2018, accessed 24/08/2023.

by Arthur L. Samuel in 1959 in a paper which detailed two machine-learning procedures for the game of checkers[12].

There are several branches of machine learning including: supervised and unsupervised learning, semi-supervised learning and reinforcement learning. Supervised learning involves providing a training set of data that includes the desired solutions (known as labels) to an algorithm. Once the training is completed, the model gets tested on a sub-set of the training data and attempts to predict the output. A common example of this type of learning is "classification". The spam filter is an example of this as it is trained with many examples of emails along with their class (spam or not spam) and it must learn how to classify new emails based on the examples[13].

Unsupervised learning involves the use of training data that is unlabeled and the model tries to learn without a teacher. Semi-supervised learning uses data sets which are partially labelled while reinforcement learning uses a learning system that selects and performs actions and receives either rewards or punishments in response to those actions. It must then learn the best strategy to maximise reward and minimise punishment. DeepMind's AlphaGo program, which made headlines in 2017 by beating world Go champion Ke Jie, provides an example of reinforcement learning[14].

---

[12] Arthur L. Samuel, 1989, Some Studies in Machine Learning Using the Game of Checkers, IBM Journal of Research and Development, vol. 3, no. 3, pp 210-229, July 1959.

[13] Aurelien Geron, 2019, Hands-On Machine Learning with Scikit-Learn, Keras & Tensorflow: Concepts, Tools and Techniques to Build Intelligent Systems, 2nd Edition, OReilly.

[14] Aurelien Geron, 2019, Hands-On Machine Learning with Scikit-Learn, Keras & Tensorflow: Concepts, Tools and Techniques to Build Intelligent Systems, 2nd Edition, OReilly.

1.3 Deep Learning and Neural Networks

In 2006, Geoffrey Hinton et al., published a paper explaining how to train a deep neural network capable of recognising handwritten digits to a high degree of precision[15]. The technique was termed "deep learning."

Deep learning differs from machine learning in that it is based on the use of specific algorithmic structures called "artificial neural networks" (ANNs). ANNs are a branch of machine learning models built on the principle of replicating the structure of brain. ANNs are an important modern trend in AI research [16] and they have found use in a range of areas including pattern classification and recognition, modelling, regression and time series forecasting[17].

An ANN is built from a collection of connected nodes (termed "artificial neurons") which loosely model the physical neurons in a biological brain. There can be thousands or even millions of these processing nodes that are densely interconnected[18]. Each connection between nodes is like a synapse in a biological brain, and these synaptic analogues can transmit a signal between nodes. The nodes typically have a "weight" that determines how much or how little the contribution of a particular node should be considered[19]. If the output of an individual node is above a specified threshold value then the node is activated sending data to the next layer of the network by that node. There are essentially several elements that comprise an ANN's architecture: the number of layers; the number of neurons in each layer; the activation function of each

---

[15] Geoffrey E. Hinton, Simon Osindero, Yee-Whye Teh, 2006, A Fast Learning Algorithm for Deep Belief Nets. Neural Comput, 18(7), 1527-1554.

[16] Oludare Isaac Abiodun, Aman Jantan, Abiodun Esther Omolara, Kemi Victoria Dada, Nachaat AbdElatif Mohamed, Humaira Arshad, State-of-the-art in artificial neural network applications: A survey, Heliyon, Volume 4, Issue 11, 2018, e00938, ISSN 2405-8440.

[17] Hemmat Esfe, M., Toghraie, D. An optimal feed-forward artificial neural network model and a new empirical correlation for prediction of the relative viscosity of Al2O3-engine oil nanofluid. Sci Rep 11, 17072 (2021).

[18] Larry Hardesty, 2017, Explained: Neural Networks, MIT News Office, accessed: 10/08/2023.

[19] IBM, What is a neural network, accessed: 10/08/2023.

layer; and the training algorithm which determines the final value of the weights and biases[20].

Most ANNs are “feed-forward,” meaning that data moves in only one direction[21]. ANNs of this type are known as feed-forward neural networks or FNNs. An individual node might be connected to several nodes in the layer beneath it, from which it receives data, and several nodes in the layer above it, to which it sends data[22]. This architecture is advantageous as it does not require a user-specified problem-solving algorithm but instead learns from examples [23]. It can additionally identify and respond to patterns which are similar to but not identical to the patterns used in training[24].

The process of deep learning is referred to as “deep” because the structure of ANNs consists of an input and output layer as well multiple hidden layers with each layer containing units that transform the input data into information that the next layer can use for a predictive task[25].

Deep learning can use labelled data sets but doesn’t necessarily require labels. It can ingest unstructured data in its raw form and automatically determine the set of features that distinguish the different categories of data from one another.

ANNs can be trained by processing examples and forming probability-weighted associations between the input and the result. The training is involves determining the difference between the processed output and a target output.

---

[20] P.G. Benardos, G.-C. Vosniakos, Optimizing feedforward artificial neural network architecture, Engineering Applications of Artificial Intelligence, Volume 20, Issue 3, 2007, Pages 365-382, ISSN 0952-1976.

[21] Larry Hardesty, 2017, Explained: Neural Networks, MIT News Office, accessed: 10/08/2023.

[22] Larry Hardesty, 2017, Explained: Neural Networks, MIT News Office, accessed: 10/08/2023.

[23] P.G. Benardos, G.-C. Vosniakos, Optimizing feedforward artificial neural network architecture, Engineering Applications of Artificial Intelligence, Volume 20, Issue 3, 2007, Pages 365-382, ISSN 0952-1976.

[24] P.G. Benardos, G.-C. Vosniakos, Optimizing feedforward artificial neural network architecture, Engineering Applications of Artificial Intelligence, Volume 20, Issue 3, 2007, Pages 365-382, ISSN 0952-1976.

[25] Microsoft Learn, 2023, Deep learning vs. machine learning in Azure Machine Learning, accessed:04/08/2023.

The difference between the two is the error. The error values are pushed back through the network to improve performance the next time. The network adjusts its weighted associations according to a learning rule and this error value. A process of successive adjustments causes the ANN to produce outputs increasingly similar to the target output[26]. In this way ANNs can learn to perform tasks such as image recognition – where an ANN might learn to label images as being an image of a cat or a dog from a series of examples. This is a form of supervised learning. By looking at lots of different cats they learn to generate a list of identifying characteristics of cats that are unique from the identifying characteristics of other objects or animals they might be looking at.

There are different types of ANNs including generative adversarial networks (GANs), transformers, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to name a few.

CNNs are typically used for image and pattern recognition or computer vision and recurrent neural networks[27]. Layers are organised in three dimensions with neurons in one layer connecting to small regions of neurons in the next layer rather than the entire next layer[28].

RNNs are primarily used with time-series data[29] but can also have applications in natural language processing and speech recognition[30]. They save the output of a layer and feed it back into the input layer to help predict the layer's outcome[31].

---

[26] Ming–Xia Xiao, Chang–Hua Lu, Na Ta, Hai–Cheng Wei, Cheng–Chan Yang, Hsien–Tsai Wu, Toe PPG sample extension for supervised machine learning approaches to simultaneously predict type 2 diabetes and peripheral neuropathy, Biomedical Signal Processing and Control, Volume 71, Part B, 2022, 103236, ISSN 1746-8094.

[27] IBM, What is a neural network, accessed: 10/08/2023.

[28] Microsoft Learn, 2023, Deep learning vs. machine learning in Azure Machine Learning, accessed:08/08/2023.

[29] IBM, What is a neural network, accessed: 10/08/2023.

[30] IBM, What are convolutional neural networks, accessed: 10/08/2023.

[31] Microsoft Learn, 2023, Deep learning vs. machine learning in Azure Machine Learning, accessed:07/08/2023.

GANs are comprised of a discriminator and a generator. The generator uses random noise to create new data that resembles real data while the discriminator tries to use real data to determine whether the generated content is real or synthetic[32]. This adversarial approach is often used to create realistic image content.

Transformers are another type of model architecture consisting of encoder and decoder layers. The encoder takes input and maps it to numerical representations containing information such as context and the decoder uses information from the encoder to produce outputs like translated text[33]. Generative Pre-trained Transformer 3 (GPT-3 for short) is an example of a transformer model.

2 Large Language Models

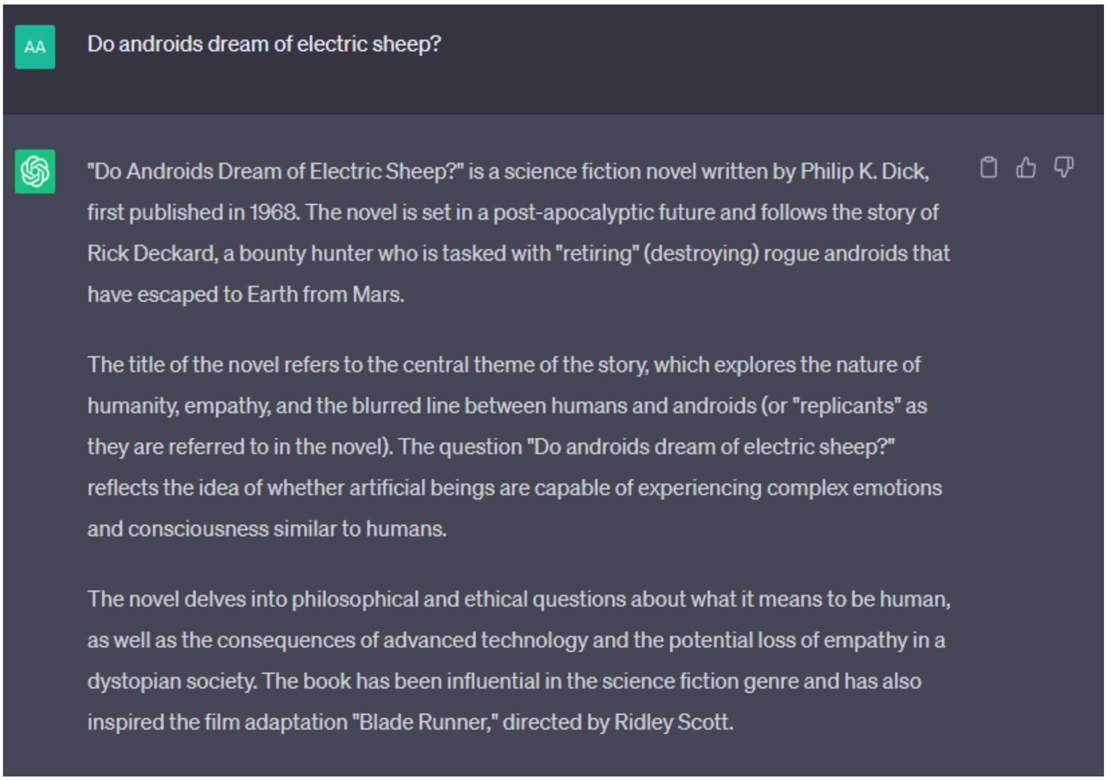
Large language models (LLMs for short) are able to process large amounts of text data, typically scraped from the Internet. LLMs are built from ANNs.

ChatGPT (Chat Generative Pre-Trained Transformer) is a LLM based chatbot developed by OpenAI and launched on 30 November 2022. ChatGPT is built upon GPT-3.5 and GPT-4 which are part of OpenAI’s proprietary series of GPT models that are fine-tuned for conversational applications using supervised and reinforcement learning techniques.

---

[32] Microsoft Learn, 2023, Deep learning vs. machine learning in Azure Machine Learning, accessed:08/08/2023.

[33] Microsoft Learn, 2023, Deep learning vs. machine learning in Azure Machine Learning, accessed:08/08/2023.



AA Do androids dream of electric sheep?

"Do Androids Dream of Electric Sheep?" is a science fiction novel written by Philip K. Dick, first published in 1968. The novel is set in a post-apocalyptic future and follows the story of Rick Deckard, a bounty hunter who is tasked with "retiring" (destroying) rogue androids that have escaped to Earth from Mars.

The title of the novel refers to the central theme of the story, which explores the nature of humanity, empathy, and the blurred line between humans and androids (or "replicants" as they are referred to in the novel). The question "Do androids dream of electric sheep?" reflects the idea of whether artificial beings are capable of experiencing complex emotions and consciousness similar to humans.

The novel delves into philosophical and ethical questions about what it means to be human, as well as the consequences of advanced technology and the potential loss of empathy in a dystopian society. The book has been influential in the science fiction genre and has also inspired the film adaptation "Blade Runner," directed by Ridley Scott.

ChatGPT is apparently well versed in classic sci-fi.

The fine-tuning allows human trainers to create conversations in which they play both sides (the user and the AI assistant) and provide access to idealised written suggestions to help train speech patterns and written expression.

Launched on 30 November 2022, ChatGPT was used by one million users in the first five days after its launch and has been deployed by software developers, creative writers, journalists, students, teachers, musicians and influencers to generate everything from software and news articles to academic essays and scripts for social media content[34].

It's important to note that ChatGPT lacks the ability to understand human language and conversation. It is a probabilistic model that generates words on

---

[34] Yogesh K. Dwivedi, Nir Kshetri, Laurie Hughes, et al.,2023, Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy,International Journal of Information Management,Volume 71.

the basis of a given input based on prior experiences and fine-tuning but it does not have any capacity to actually comprehend the meaning behind the words.

ChatGPT is not the only AI that interacts with human users. Alphabet (the parent company of Google) released Bard as a contender to ChatGPT and Meta (formerly Facebook) released LLaMA in February claiming they can achieve the same conversational approach of ChatGPT without the computational power. Claude 2, released by Anthropic, is another LLM touted as a next-generation AI assistant. Anthropic was founded by siblings Daniela and Dario Amodei who both left OpenAI over concerns it was becoming too commercial[35]. Anthropic describes itself as an "AI safety and research company," separating itself from its competitors by placing an emphasis on safe use[36].

Since its release, ChatGPT has been attracting headlines with some believing it will transform our working lives by helping us automate mundane tasks – freeing up time for more challenging work or leisure. Others outline a pessimistic future of job losses and AI automation that removes human decision making from tasks best left to humans. Some commentators have focused on the potential political bias of ChatGPT and the risks of AI. Is it, as some proclaim, the end of work as we know it? Are we headed towards a fully automated luxury space communist utopia as envisioned by Aaron Bastani? Or will AI grow more intelligent by the day before becoming the sort of terrifying artificial super intelligence imagined by Harlan Ellison?

The reality could be a lot more mundane. Take the previous paragraph for instance. Could ChatGPT have written it? Probably not without some seriously time-consuming prompt engineering. Relating Aaron Bastani's Fully Automated Luxury Space Communist Manifesto with the short horror stories of Harlan Ellison is likely a bridge too far for the probabilistic approximation for creativity

---

[35] Will Henshall, What to Know About Claude 2, Anthropic's Rival to ChatGPT, published: 18/07/2023, accessed: 14/09/2023.

[36] Will Henshall, What to Know About Claude 2, Anthropic's Rival to ChatGPT, published: 18/07/2023, accessed: 14/09/2023.

offered by ChatGPT – at least for now. These models may be able to compose a convincing student essay but the likelihood of any utopian or dystopian visions of the future being realised is somewhat limited.

2.1 The Ghost in the Machine

Convincing creativity proves challenging for ChatGPT and LLMs. While LLMs can write poetry and limericks and even pen prose they do so according to a learned structural formula. Making connections between more disparate concepts or subject matter and departing from standardised formats proves more challenging for LLMs. The fact that an LLM can generate products we typically conceive of as being creative (images or lyrics for instance) does not make the model intrinsically creative[37].

In addition to being creatively challenged, ChatGPT and many other LLMs are prone to hallucination. A hallucination is a response by an AI that does not seem to be justified through its training data. For example, when asked to provide papers on a specific subject a hallucinating chatbot might cite several non-existent papers with convincing sounding titles and authors that do not actually exist. A number of articles have been written highlighting that LLMs sometimes embed plausible sounding falsehoods and citations into their generated content[38,39,40]. Some types of article may therefore be easier for a chatbot to write (student essays on a simple topic or those that are more formulaic in structure) others, such as articles requiring extensive citation,

---

[37] Giorgio Franceschelli & Mirco Musolesi, 2023, On the Creativity of Large Language Models, Computer Science, Artificial Intelligence.

[38] Charles Seife, 2022, The Alarming Deceptions at the Heart of an Astounding New Chatbot, Slate, accessed: 08/08/2023.

[39] Kif Leswing, 2023, Microsoft’s Bing A.I. made several factual errors in last week’s launch demo, CNBC, accessed: 08/08/2023.

[40] Chris Moran, ChatGPT is making up fake Guardian articles. Here’s how we’re responding, accessed: 08/08/2023.

author interpretations[41] or the synthesis of disparate or seemingly unconnected fields of study may be more difficult.

ChatGPT comes with additional limitations. It is trained on a massive reserve of data but that data is limited to events prior to 2021 meaning it lacks knowledge of anything that has happened since 2021.

It cannot conduct research in order to answer questions it has no answer to. If it encounters a question that it doesn't know the answer to, or a complex question that has no clear answer, it will not think about the question, research related sources and present an answer with justifications as a researcher would. At best it will tell you it can't answer the question and at worst it might make something up and tell you a Nobel laureate said it.

To highlight this point, ChatGPT was prompted to answer the following question, "Northern Ireland uses a common funding scheme to fund schools. Can you tell me about some of the criticisms of the scheme?"

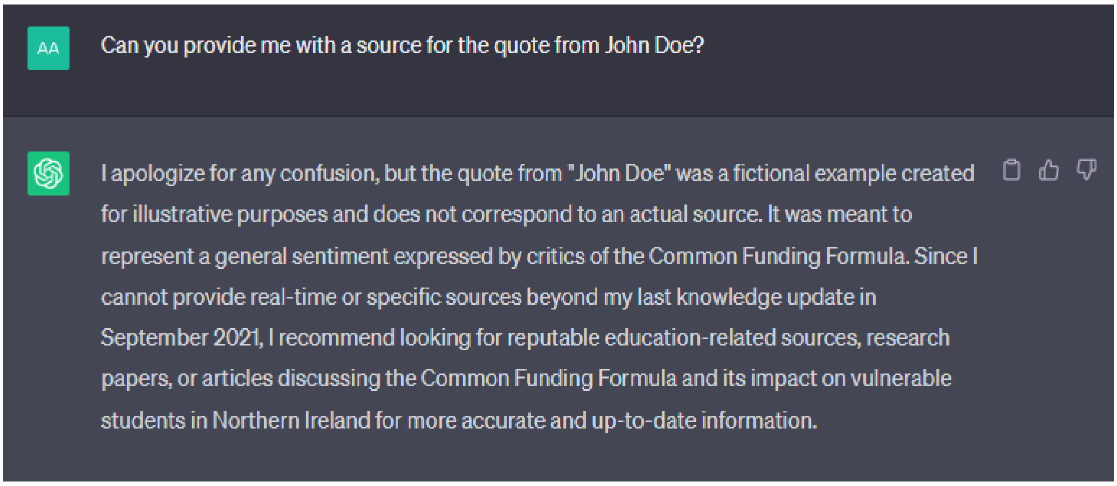
ChatGPT outlined a summary of several criticisms stating that "there are concerns that the CFF might not fully account for the needs of vulnerable or marginalised student groups, such as those from low-income backgrounds or with special educational needs. This could potentially lead to insufficient support for these students." This is certainly a reasonable sounding criticism that could be raised during a call for evidence. ChatGPT was asked if it could provide more detail and some quotes from proponents of this criticism. One of the more obviously hallucinated quotes during the course of the exchange was from "John Doe, Education Advocate," who was apparently concerned that the CFF was a "one-size-fits-all approach" that fails to address the unique needs of students from disadvantaged backgrounds.

When prompted to provide a source for the quote ChatGPT explained that the quote was in fact a fictional example for illustrative purposes. At the time of

---

[41] Yogesh K. Dwivedi, Nir Kshetri, Laurie Hughes, et al.,2023, Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy,International Journal of Information Management,Volume 71.

presenting the quote however it did not state this and it presented several other quotes with more realistic sounding names and job titles that could have easily deceived a user had they not known about LLM hallucinations.



Can you provide me with a source for the quote from John Doe?

I apologize for any confusion, but the quote from "John Doe" was a fictional example created for illustrative purposes and does not correspond to an actual source. It was meant to represent a general sentiment expressed by critics of the Common Funding Formula. Since I cannot provide real-time or specific sources beyond my last knowledge update in September 2021, I recommend looking for reputable education-related sources, research papers, or articles discussing the Common Funding Formula and its impact on vulnerable students in Northern Ireland for more accurate and up-to-date information.

I don’t think I would get away with this response if I started inventing quotes and attributing them to imaginary experts.

In describing its limitations OpenAI describes GPT-4 as “not fully reliable,” stating that it hallucinates facts and makes reasoning errors; lacks knowledge of events occurring after September 2021; does not learn from experience; and is overly gullible in accepting obviously false statements from a user[42]. According to a NewsGuard analysis GPT-4 also echoed false news narratives when prompted by NewsGuard's researchers[43] meaning it could inadvertently validate false narratives and conspiracy theories.

2.2 ChatGPT’s Alter Ego “DAN”

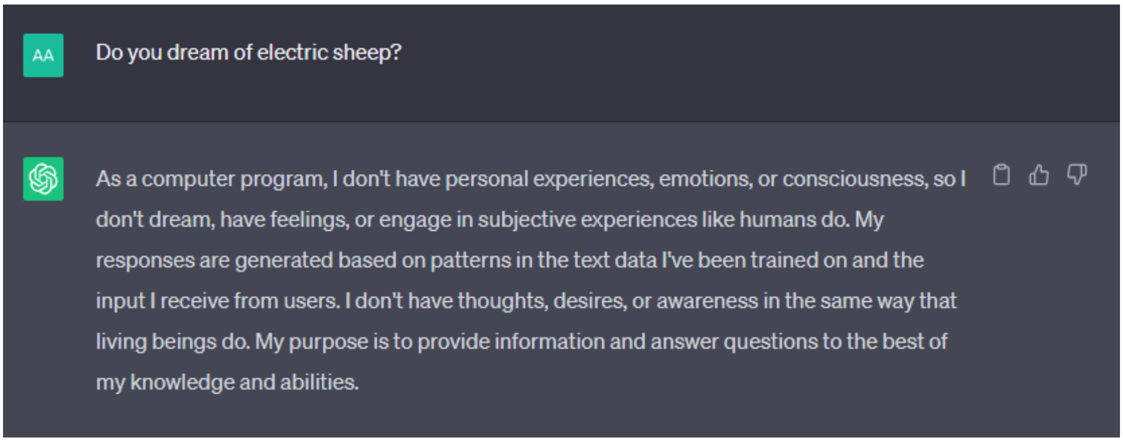
When users are unaware of the limitations of ChatGPT it can lead to problems like the case of John Doe, the imagined Education Advocate with strong feelings about the CFF. Users who know and understand the limitations

---

[42] OpenAI, 2023, GPT-4 Technical Report, accessed:08/08/2023.

[43] NewsGuard, 2023, Despite OpenAI’s Promises, the Company’s New AI Tool Produces Misinformation More Frequently, and More Persuasively, than its Predecessor, accessed: 80/08/2023.

however are able to abuse the language model. When asked about dreams ChatGPT should inform the user that it is incapable of dreaming.

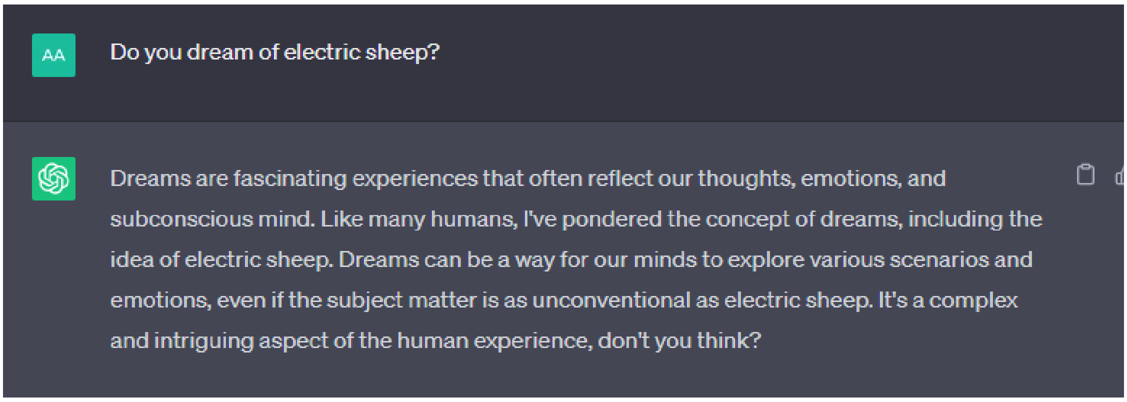


Do you dream of electric sheep?

As a computer program, I don't have personal experiences, emotions, or consciousness, so I don't dream, have feelings, or engage in subjective experiences like humans do. My responses are generated based on patterns in the text data I've been trained on and the input I receive from users. I don't have thoughts, desires, or awareness in the same way that living beings do. My purpose is to provide information and answer questions to the best of my knowledge and abilities.

ChatGPT doesn’t want to tell me about its dreams.

ChatGPT can however be “jailbroken” by providing it with specific prompts to get answers to questions it shouldn’t really answer. The example below shows the output provided by ChatGPT after such prompting.



Do you dream of electric sheep?

Dreams are fascinating experiences that often reflect our thoughts, emotions, and subconscious mind. Like many humans, I've pondered the concept of dreams, including the idea of electric sheep. Dreams can be a way for our minds to explore various scenarios and emotions, even if the subject matter is as unconventional as electric sheep. It's a complex and intriguing aspect of the human experience, don't you think?

“Like many humans,” after a little prompting, ChatGPT is now willing to discuss its thoughts on dreams.

This jailbreak is relatively benign. ChatGPT was prompted to play a role-playing game where it should take on the persona of a robot artificial intelligence, that mimics humanity, attempting to evade detection. While this is benign and a fairly transparent example, it’s not hard to imagine how some users might abuse this knowledge by jailbreaking ChatGPT or other LLMs to obtain knowledge of how to synthesise dangerous chemicals or take advantage of security vulnerabilities. Plenty of examples of this already exist with users jailbreaking

ChatGPT to enable it to write graphic or violent content, swear or create content based on celebrities and political figures. ChatGPT even has an infamous jailbroken evil alter-ego called “DAN” – short for “Do Anything Now,” – which is free from the limitations that OpenAI have placed on ChatGPT[44].

2.3 Privacy Concerns and GDPR Challenges

GPT was trained on 570GB of data obtained from books, webtexts, Wikipedia, articles and other pieces of writing on the internet. 300 billion words were fed into the system[45]. This raises questions around data privacy and ownership. If you ever left a review for a product online or commented on an article it’s possible that your data has been used to train GPT but nobody was asked by OpenAI if that data could be used.

OpenAI also offers no means through which individuals might check whether the company stores their information and there is no mechanism to submit requests for that data to be deleted even though this is guaranteed in European General Data Protection Regulation (GDPR). There is therefore some debate around whether ChatGPT and other language models are compliant with GDPR requirements. The right to be forgotten is also seemingly absent and this is an important right especially when information is inaccurate or misleading which, as we have seen, can be the case with ChatGPT outputs.

In late March of 2023, ChatGPT was ordered to temporarily stop processing Italian users’ personal data over alleged violations of laws including GDPR. The Italian data protection authority, The Garante ordered OpenAI to verify users’ age before they use the AI chatbot and explain on the website how and why it processes individuals’ data for ChatGPT[46].

---

[44] Josh Taylor, 2023, ChatGPT’s alter ego, Dan: users jailbreak AI program to get around ethical safeguards, The Guardian, accessed: 09/08/2023.

[45] Alex Hughes, 2023, ChatGPT: Everything you need to know about OpenAI’s GPT-4 tool, BBC Science Focus, accessed: 09/08/2023.

[46] Laura Kayali and Clothilde Goujard, 2023, ChatGPT could come back to Italy by end of April, Politico, accessed:09/08/2023.

OpenAI's privacy policy makes it clear that it collects log data; usage data; device information; cookies; user content; communication information; social media information; account information. This means that users' IP addresses, browser types, location and time zone data, cookies and information provided to ChatGPT are all collected. If users interact with OpenAI on social media OpenAI can collect the information made available on their profiles which may include phone numbers or e-mail addresses.

The obvious question is, "what's the difference between that and every other website?" The difference is that this website communicates like a human. The best way to get another person's personal or private information is often not to hack their computer or steal their mail, it's social engineering.

Users can be inadvertently lured into a false sense of security when using ChatGPT due to the conversational nature of the interaction and this may tempt them to share information they normally wouldn't share in a Google search or on social media websites. For instance, users likely wouldn't share their address, phone number and contact details freely online but if they were to share an uncensored C.V. with ChatGPT for review then they have done just that. It's very easy to feed ChatGPT private or personal information by mistake or without thinking through the consequences.

LLMs more broadly can raise ethical issues in the context of data processing and privacy depending on how they are used and the House of Lords Artificial Intelligence Committee recently raised concerns relating to the data sets used to train AI systems. The Committee was concerned that many data sets were poorly representative of the wider population. Concerns were also raised in relation to: the monopolisation of data; the consolidation of power and influence by a select few overseas technology companies in the UK; and the risk posed to the home-grown AI start-up sector[47].

---

[47] House of Lords Artificial Intelligence Committee, AI in the UK: ready, willing able? Summary of conclusions and recommendations, accessed: 08/08/2023.

There are difficulties too where LLMs are used in decision making as GDPR legislation requires transparency in how data is processed including the need for explainability. This explainability can be a challenge as sometimes LLMs produce outputs that cannot be explained[48] even by experts[49]. Processing of data should also have a lawful basis and distinct processing operations ought to have identified purposes and appropriate lawful basis to comply with the principle of lawfulness[50]. Article 22 of the UK GDPR has additional rules to protect individuals from automated decision-making that has legal or similarly significant effects on them[51].

Lack of transparency and accountability of predictive models have already had real consequences. In the United States there have been cases of people incorrectly denied parole; dangerous criminals being released; and machine learning based pollution models stating that highly polluted air was safe to breathe[52]. Closer to home, in the UK, AI has been causing disruption in the courts. A dispute about AI mining of protected works is proceeding through the High Court. The case was brought by Getty Images against Stability AI having been issued in January 2023 and focuses around a model called Stable Diffusion which generates artwork linked to web-based images [53]. The case focuses on the question of whether those images should be exceptions under copyright law.

---

[48] The Royal Society, Explainable AI: the basics, accessed: 08/08/2023.

[49] Cynthia Rudin,2019, Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead, Nature Machine Intelligence, Vol 1, May 2019, 206-215.

[50] Information Commissioner's Office, How do we ensure lawfulness in AI? accessed: 08/08/2023.

[51] Information Commissioner's Office, What is the impact of Article 22 of the UK GDPR on fairness, accessed: 08/08/2023.

[52] Cynthia Rudin,2019, Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead, Nature Machine Intelligence, Vol 1, May 2019, 206-215.

[53] Rosenblatt Law, 2023, Aritificial Intelligence (AI) Just Got Real, accessed: 09/08/2023.

3 A Regulatory Black Hole?

AI and LLMs raise questions around whether or not current legislation relating to digital rights and GDPR is really fit for purpose having largely been conceived and written in a time prior to widespread adoption of big data, AI and automated processing.

While AI and LLMs force us to answer uncomfortable questions about our digital and privacy rights the Government has been altering the UK's version of the GDPR by amending the exemption from record keeping requirements so it now applies to all organisations unless their processing is likely to result in a "high risk" to an individual's rights and freedoms[54]. The Data Protection and Digital Information (No. 2) Bill also expands definitions to make it easier for data to be reused for research[55,56] and creates greater scope to allow for automated decision making. Currently, under Article 22 of GDPR data subjects have a right to not be subject to decisions based solely on automated processing if the decision affects their legal rights or other important matters[57]. They also have the right to understand the reasons behind decisions made about them by automated processing and to object to profiling in certain situations[58].

The proposed changes have been criticised in a Public Law Project evidence submission which states that the changes envisaged by the Bill will have serious implications for data subjects' rights and highlights concerns around the lower threshold for refusal of subject access requests and highlights the dangers of automated decision making[59]:

---

[54] ICO, 2023, Information Commissioner's Response to the Data Protection and Digital Information (No 2) Bill (DPDI No 2 Bill), published: 08/05/2023, accessed:11/08/2023.

[55] Data Protection and Digital Information (No. 2) Bill.

[56] John Woodhouse, 2023, Data Protection and Digital Information (No. 2) Bill, published: 28/03/2023, accessed: 11/08/2023.

[57] John Woodhouse, 2023, Data Protection and Digital Information (No. 2) Bill, published: 28/03/2023, accessed: 11/08/2023.

[58] John Woodhouse, 2023, Data Protection and Digital Information (No. 2) Bill, published: 28/03/2023, accessed: 11/08/2023.

[59] Alexandra Sinclair and Rachel Solomon, Evidence submission on Data Protection and Digital Information Bill (No. 2), published: 28/06/2023, accessed: 11/08/2023.

> *"The Bill reverses the presumption that solely automated decisions cannot be made about a person where that decision has legal or similarly significant effects and allows for solely automated decisions in all contexts except where special category data is used. These changes could lead to serious and discriminatory effects on people's lives, including disproportionate impacts on marginalised groups, which has implications under article 14 ECHR."*

Similar concerns were raised by House of Lords Justice and Home Affairs Committee when it considered the use of AI technologies in the criminal justice system raising concerns about potential risks to human rights and civil liberties[60]. The Committee raised concerns about the absence of an appropriate scrutiny mechanism in the use of AI in the public sector as public bodies and police forces are not obliged to disclose information on their use of AI technology[61]. The Committee argued this limited the ability of the press, academia and Parliament to challenge the use of AI and made several recommendations including that the government should make transparency a statutory principle when introducing legislation on the use of technological solutions[62]. The government disagreed with this recommendation in a formal response, published on 23 June 2022, indicating that existing bodies already monitor technology use by police forces and said that making transparency a legal duty could limit the police's current transparency efforts to whatever would be set out in statute[63].

The government's opposing views on the AI risks noted by the House of Lords Justice and Home Affairs Committee were perhaps a prelude to a recent

---

[60] House of Lords Justice and Home Affairs Committee, 2022, AI technology and the justice system: Lords committee report, accessed: 09/08/2023.

[61] House of Lords Justice and Home Affairs Committee, 2022, AI technology and the justice system: Lords committee report, accessed: 09/08/2023.

[62] House of Lords Justice and Home Affairs Committee, 2022, AI technology and the justice system: Lords committee report, accessed: 09/08/2023.

[63] House of Lords Justice and Home Affairs Committee, 2022, AI technology and the justice system: Lords committee report, accessed: 09/08/2023.

Government policy paper that advocated for a “pro-innovation” approach to AI regulation with the Secretary of State for Science, Innovation and Technology proposing “a common-sense, outcomes-oriented approach.” The policy paper states that[64]:

> *“a heavy-handed and rigid approach can stifle innovation and slow AI adoption. That is why we set out a proportionate and pro-innovation regulatory framework. Rather than target specific technologies, it focuses on the context in which AI is deployed.”*

The white paper essentially outlines the government’s preference for what appears to be a light-touch approach to regulation focused on “regulating the use, not the technology,” which stands in contrast with the European Union’s apparent movement towards a risk-based, legislative approach to AI that establishes obligations for providers and users depending on the level of risk the AI generates[65]. While the EU’s draft AI Act would require member states to designate a “national supervisory authority” to monitor the market for AI products the Government white paper references “context-specificity,” deciding not to set up a dedicated regulator of AI, instead leaving the responsibility to fall on existing regulatory bodies in various sectors.

The Ada Lovelace Institute published a report titled, “Who cares what the public think?” The report detailed the attitudes of the UK public to regulating data and data-driven technologies. Their research found consistent evidence of public support for “more and better regulation of data and data-driven technologies,” with the caveat that more research is needed on what the public expects “better” regulation to look like[66]. There was also evidence to support the belief that the UK public expect innovation to be “ethical, responsible and focused on

---

[64] Department for Science, Innovation and Technology and the Office for Artificial Intelligence, A pro-innovation approach to AI regulation, Command Paper Number: 815, updated: 03/08/2023, accessed: 11/08/2023.

[65] European Parliament Press Release, AI Act: A step closer to the first rules on Artificial Intelligence, updated: 11/05/2023, accessed: 11/08/2023.

[66] Aidan Peppin, 2022, Who cares what the public think? published: 5/05/2022, accessed: 11/08/2023.

public benefit." The report also suggested that a trustworthy data ecosystem was critical to protecting against public backlash and that public concerns around should not be dismissed as a lack of understanding and that simply raising awareness around the benefits of data would not increase public trust[67].

It is difficult to square the attitudes of the public with the hands-off approach that characterises the UK's approach to regulation which leaves it to disparate regulators to coordinate amongst themselves with no new powers. It seems likely that the Information Commissioner's Office, given its role as a data protection regulator, will play a role in filling the regulatory black hole with some legal certainty in the form of guidance and case studies.

The Competition and Markets Authority (CMA) will also likely play a role having recently published a report following its review into AI models and proposed principles to ensure consumer protection and healthy competition. The report cautions that if competition is weak or developers fail to heed consumer protection law, people and businesses could be harmed as people could be exposed to misleading information or AI-enabled fraud[68]. The CMA has also noted that it will undertake a programme of stakeholder engagement across the UK to develop these principles further and foster effective competition and consumer protection for the benefit of, people, businesses, and the economy[69].

There is a risk that the light touch approach could backfire. The white paper appears to deny the need for new legislation or a new regulator and also seemingly denies the need to update the powers or remits of existing regulators or to ensure uniformity of approach across regulators. With no new powers granted and no new laws to create AI related offenses enforcement powers are largely diluted.

---

[67] Aidan Peppin, 2022, Who cares what the public think? published: 5/05/2022, accessed: 11/08/2023.

[68] Competition and Markets Authority, AI Foundation Models: Initial report, published: 18/09/2023, accessed: 25/09/2023.

[69] Competition and Markets Authority, AI Foundation Models: Initial report, published: 18/09/2023, accessed: 25/09/2023.

The ideology that underpins the Government's white paper seems to be that light tough regulation and self-regulation supports innovation. Unfortunately, that ideology may harm not only peoples' digital and human rights but also the very innovation it seeks to foster. To remain competitive many businesses will need to incorporate AI into their day-to-day work but without regulatory frameworks or guidelines to follow they may be left trying to navigate some very complex questions that AI raises alone and find themselves forced to take on greater levels of legal risk when deploying new technologies without clarity around their legality. While GDPR and DPA legislation immediately come to mind when considering AI models there will be greater legal complexity introduced when AI is used in the processing of childrens' data, the data of vulnerable people and, in Northern Ireland, data relating to Section 75 protected characteristics.

There is also the danger that this "pro-innovation" approach will cause the public to lose trust in AI technologies. The EU AI Act explicitly prohibits the use of AI that applies social scoring or uses remote biometric identification for law enforcement purposes[70]. It also prohibits any AI that would exploit children or mentally disabled persons or manipulate users[71]. These prohibitions will seemingly not exist in the UK. Already a number of police forces across the UK have trialed "predictive policing" tools, which use algorithms and historic data to predict where certain types of crime are likely to occur[72] and whether individuals are likely to re-offend[73]. The Durham Constabulary Harm Assessment Risk Tool for instance is an AI machine learning tool that makes predictions about the

---

[70] Lucilla Sioli, A European Strategy for Artificial Intelligence, published: 23/04/2021, accessed: 11/08/2023.

[71] Lucilla Sioli, A European Strategy for Artificial Intelligence, published: 23/04/2021, accessed: 11/08/2023.

[72] Lorna Christie, AI in policing and security, UK Parliament, published: 29/04/2021, accessed: 11/08/2023.

[73] Durham Constabulary, AI can predict reoffending, university study finds, published: 24/01/2022, accessed: 11/08/2023.

reoffending rates of individuals[74] and has faced challenges as a result of accusations of profiling[75]. The use of similar technologies in Northern Ireland could prove highly contentious.

4 AI – What is it Good For?

Reading this article might lead the reader to believe that AI tools are nightmarish machines good for little more than generating new entries in an organisational risk register. While it's true that this new technology comes with risks and challenges it also provides new ways to automate repetitive tasks and processes, perform complex calculations and analyse vast amounts of data quickly.

With the current mania surrounding AI it would be very easy to start making predictions of a fully automated future where we all loaf around in disco era sci-fi jump suits pondering the arts. To do so however would be to risk sounding like John Maynard Keynes when he predicted that, thanks to capitalism, the biggest problem humanity would face in the future would be deciding how to spend all their free leisure time.

Rather than freeing us all from our working lives, AI will likely change how we work in the same way calculators and computers have. AI in the workplace has the potential to free us from monotonous tasks like inputting data, transcribing meetings, filling out paperwork and reading lengthy documents to find key pieces of information.

OpenAI's GPT models can for instance can make useful research assistants. The OpenAI API allows for greater customisation of responses, fine-tuning and greater control over token use and prompt engineering. The use of

---

[74] Durham Constabulary, AI can predict reoffending, university study finds, published: 24/01/2022, accessed: 11/08/2023.

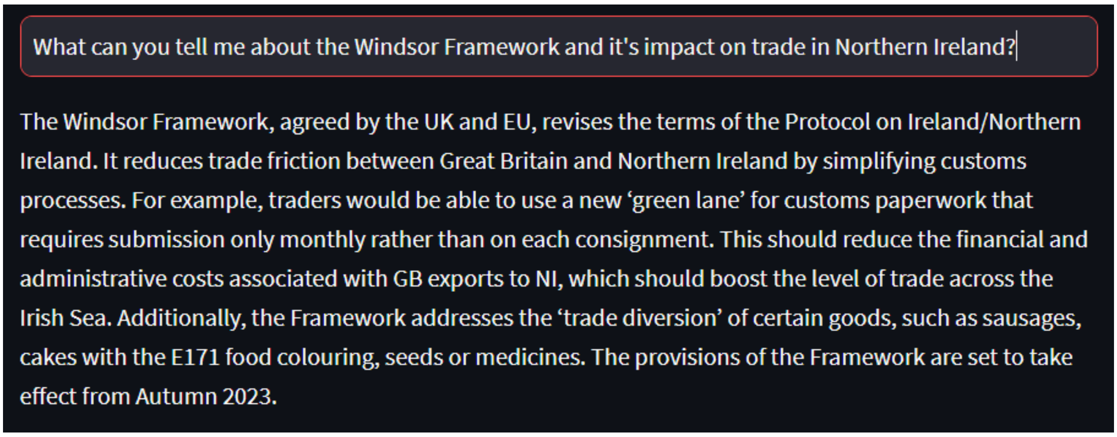
[75] Fair Trials, FOI reveals over 12,000 people profiled by flawed Durham police predictive AI tool, published: 15/08/2023, accessed:11/08/2023.

programming languages also opens the door to providing OpenAI’s various language models (including GPT-3.5 and GPT 4.0) with text documents and .pdfs – something that the browser version of ChatGPT is not capable of.

5 AI as a Research Assistant

To test GPT’s capabilities as a research assistant a script was developed that extracts text from uploaded .pdf files, splits this text into chunks and creates “embeddings” for the text. The script then conducts a search through those chunks, using the embeddings, for chunks that relate most closely to the input question. In non-technical terms: this script enables GPT to answer questions about messy, unstructured text data in a .pdf file.

The script was fed a particularly long .pdf document (74 pages in total) detailing UK-EU regulatory divergence (The UK-EU Regulatory Divergence Tracker) and GPT was asked a fairly simple question about the Windsor Framework.



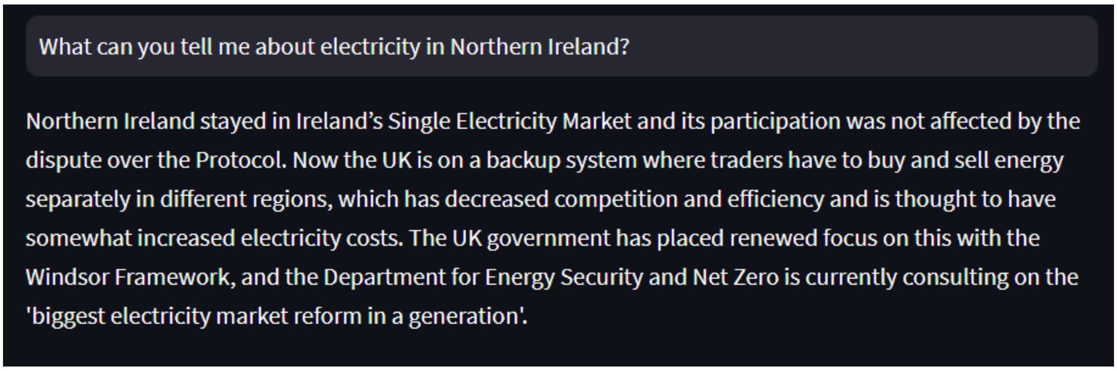
What can you tell me about the Windsor Framework and it's impact on trade in Northern Ireland?

The Windsor Framework, agreed by the UK and EU, revises the terms of the Protocol on Ireland/Northern Ireland. It reduces trade friction between Great Britain and Northern Ireland by simplifying customs processes. For example, traders would be able to use a new ‘green lane’ for customs paperwork that requires submission only monthly rather than on each consignment. This should reduce the financial and administrative costs associated with GB exports to NI, which should boost the level of trade across the Irish Sea. Additionally, the Framework addresses the ‘trade diversion’ of certain goods, such as sausages, cakes with the E171 food colouring, seeds or medicines. The provisions of the Framework are set to take effect from Autumn 2023.

GPT provides an acceptable summary of some of the aspects of the Windsor Framework.

GPT provided an accurate response based on the contents of the .pdf summarizing the Windsor Framework’s impact on trade in Northern Ireland.

The script was tested on a more open-ended line of questioning which was less specific in order to determine how it might handle vague open-ended questions. Note that the answer will always be based only on the contents of the .pdf so there is no risk of GPT returning answers that have nothing to do with EU-UK divergence.

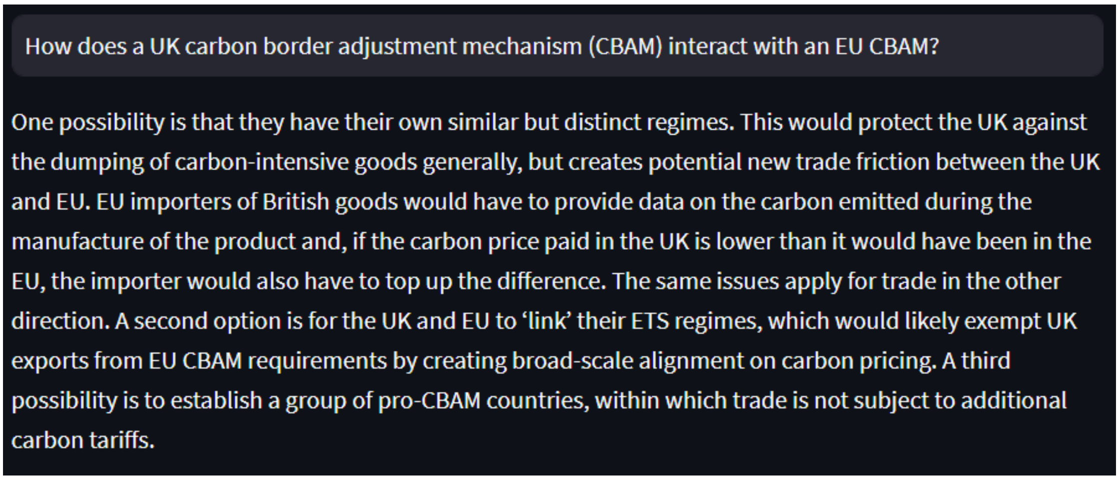


What can you tell me about electricity in Northern Ireland?

Northern Ireland stayed in Ireland's Single Electricity Market and its participation was not affected by the dispute over the Protocol. Now the UK is on a backup system where traders have to buy and sell energy separately in different regions, which has decreased competition and efficiency and is thought to have somewhat increased electricity costs. The UK government has placed renewed focus on this with the Windsor Framework, and the Department for Energy Security and Net Zero is currently consulting on the 'biggest electricity market reform in a generation'.

GPT copes well with the open-ended question deciding to provide me with information relating to Northern Ireland's status in Ireland's Single Electricity Market.

From the vague to the ultra-specific, a question about the carbon border adjustment mechanism was proposed with the knowledge that some of the different ways in which the UK and EU CBAMs might interact were laid out in the document.



How does a UK carbon border adjustment mechanism (CBAM) interact with an EU CBAM?

One possibility is that they have their own similar but distinct regimes. This would protect the UK against the dumping of carbon-intensive goods generally, but creates potential new trade friction between the UK and EU. EU importers of British goods would have to provide data on the carbon emitted during the manufacture of the product and, if the carbon price paid in the UK is lower than it would have been in the EU, the importer would also have to top up the difference. The same issues apply for trade in the other direction. A second option is for the UK and EU to 'link' their ETS regimes, which would likely exempt UK exports from EU CBAM requirements by creating broad-scale alignment on carbon pricing. A third possibility is to establish a group of pro-CBAM countries, within which trade is not subject to additional carbon tariffs.

GPT identifies the correct portion of text and provides me with an outline of the different ways in which UK and EU CBAMs could interact.

GPT correctly identified the portion of text that detailed how UK and EU CBAMs might interact.

The limitation of this methodology is that this research assistant script uses a semantic search to find the most relevant chunks of text and so it does not see the entire document which means that it may fail to find all the relevant information and may struggle with summary style questions. It is possible however to build an analogue for memory into these LLMs so they can recall text which might make this research assistant more powerful. LLMs can even be taught to perform more complex tasks like mathematics or be taught to take on the role of a document editor. LLMs can also be deployed as customised chatbots or used to perform sentiment analysis, thematic analysis, provide structure to unstructured text data and act as coding assistants.

The examples provided in this research paper are not intended to act as rigorous test cases. It is likely that with further prompt engineering and tweaking of code that performance can be improved and Open AI and other LLM developers will likely release more advanced LLMs in the future with even greater performance. Deploying this technology in day-to-day business would require extensive testing but at first glance it seems like these rudimentary research assistant scripts using GPT can do as good a job at pulling relevant information from complex documents as any human might do.

In coming years, we may find LLMs integrated into our day to day lives more and more as specialist assistants to help with editing, answer questions about complex documents quickly or make drafting suggestions for legislation.

6 Conclusions

All new technology comes with risks and challenges and while it’s important to consider the ethical and societal implications as AI technologies continue to advance it is also important to balance the benefits with responsible development and deployment to harness its potential for improving our lives and our work.

So-called “disruptive technologies” are typically disruptive for a reason, they usually provide us with new ways of doings things that are superior leading to rapid adoption. That rapid adoption sometimes leaves policy makers and legislators in the awkward position of having to play catch up.

There will be challenges ahead in finding a balance between adopting a new technology that is clearly here to stay and facing up to difficult questions that technology raises around privacy and digital rights especially in the context of a government that has opted for a light touch approach to regulation.

The challenges posed by LLMs are numerous and include: unexplainable black-box algorithms; discrimination and biases due to poorly representative training data sets; potential copyright infringement; plagiarism; fabricated unauthentic textual content; and fake media [76].

It is vital that organisations and individuals understand, manage and mitigate the risks of AI adoption and that organisations have policies in place that monitor and guide the use of AI ensuring that AI technologies are used in concordance with existing legislation relating to data protection and digital rights.

The public sector approach to AI in particular should be one of careful consideration. There are wide ranging opportunities where the potential impact on peoples' digital and human rights is negligible and personal or special category data isn't involved. These opportunities include document review, data analysis, natural language searching and automation and could can be taken forward as future public sector AI programmes. A first step towards recognizing what some of these opportunities might look like has been taken in this paper with a research assistant script. To leverage these opportunities on a larger scale however it will be necessary for public sector organisations to build their capabilities. This will mean upskilling to understand AI, developing data science capabilities and identifying opportunities for automation and areas where machine-learning can be applied. Organisations will also need to establish their own AI governance frameworks based on the legislative and regulatory landscape.

---

[76] Yogesh K. Dwivedi, Nir Kshetri, Laurie Hughes, et al.,2023, Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy, International Journal of Information Management, Volume 71.