Artificial Intelligence (AI) Futures: India-UK Collaborations Emerging from the 4th Royal Society Yusuf Hamied Workshop[1]

Yogesh K. Dwivedi
Digital Futures for Sustainable Business & Society Research Group, School of Management, Swansea University, Bay Campus, Fabian Bay, Swansea, Wales, UK

Symbiosis Institute of Business Management, Pune & Symbiosis International (Deemed University), Pune, India
Email: y.k.dwivedi@swansea.ac.uk

Laurie Hughes
School of Business and Law, Edith Cowan University, Joondalup WA, Australia,
Email: david.hughes@ecu.edu.au

Harshad K.D.H. Bhadeshia
Department of Materials Science and Metallurgy, University of Cambridge, 27 Charles Babbage Road, Cambridge CB3 0FS, UK

School of Engineering and Materials Science, Queen Mary University of London, Mile End Road, London E1 4NS, UK
Email: h.bhadeshia@qmul.ac.uk

Sophia Ananiadou
Department of Computer Science, National Centre for Text Mining, The University of Manchester, UK
Email: Sophia.Ananiadou@manchester.ac.uk

Anthony G. Cohn
School of Computing, University of Leeds, Leeds, UK
Alan Turing Institute, UK
Department of Computer Science and Technology, Tongji University, China
Email: A.G.Cohn@leeds.ac.uk

Jacqueline M. Cole
Cavendish Laboratory, University of Cambridge, J J Thomson Avenue, Cambridge, CB3 0HE, UK

ISIS Neutron and Muon Source, STFC Rutherford Appleton Laboratory, Harwell Science and Innovation Campus, Didcot, OX11 0QX. UK
Email: jmc61@cam.ac.uk

[1]**Disclaimer**: This editorial was developed in collaboration with authors who participated in the 4th Royal Society Yusuf Hamied Workshop. However, it is important to clarify that the Royal Society did not play a role in the development, endorsement, or publication of this editorial. The views, thoughts, and opinions expressed herein belong solely to the authors and contributors of this editorial and do not reflect the official stance or position of the Royal Society.

Gareth Conduit
Cavendish Laboratory, University of Cambridge, J J Thomson Avenue, Cambridge, CB3 0HE, UK
Email: gjc29@cam.ac.uk

Maunendra Sankar Desarkar
Department of CSE and Department of AI, Indian Institute of Technology Hyderabad, India
Email: maunendra@cse.iith.ac.in

Xinwei Wang
School of Engineering and Materials Science, Queen Mary University of London, E1 4NS London, UK
Email: xinwei.wang@qmul.ac.uk

Highlights

1. Innovation through AI can make a transformative impact on industry and society
2. Trustworthy AI is critical for increasing levels of adoption
3. Evaluating the performance of LLMs to include common sense reasoning
4. Federated learning algorithms can improve the performance of AI models

Abstract

"Artificial Intelligence" in all its forms has emerged as a transformative technology that is in the process of reshaping many aspects of industry and wider society at a global level. It has evolved from a concept to a technology that is driving innovation, transforming productivity and disrupting existing business models across numerous sectors. The industrial and societal impact of AI is profound and multifaceted, offering opportunities for growth, efficiency, and improved healthcare, but also raising ethical and societal challenges as the method is integrated into many aspects of human life and work. This editorial is developed by contributors of the 4th Royal Society Yusef Hamied Workshop on Artificial Intelligence, designed to enhance collaboration between Indian and the UK scientists and to explore future research opportunities. The insights shared at the workshop are shared here.

Keywords: Artificial Intelligence; ChatGPT; Generative AI; Gen AI; Large Language Models; Technological Disruption, uncertainties

1. Introduction

In the context of technological advancement and adoption, “Artificial Intelligence” (AI) or “machine learning”, has exhibited a meteoric rise in a relatively short space of time as both industry and society grapples with the impact of its utility. The AI and associated methods form some of the basis of self-driving cars, search engines, robotics as well as established software tools and products, thus affecting the way that people live and work. It has the potential to transform industry as well as nurture new human-machine interactions and practice at a global level (Dwivedi *et al*., 2021a). Research from Goldman Sachs reminds us that 60% of current employees are in occupations that did not exist in 1940, attributing more than 85% of growth over the last eight decades is due to the impact of technology. The integration of machine learning with natural language processing within a variety of devices and systems in businesses and society, could drive an increase in global GDP and productivity (Goldman Sachs, 2023).

AI offers opportunities for innovation, efficiency, and the potential for improvements in the quality of life. Medical images can now be diagnosed in real time with machine learning, increasing both the speed and accuracy of diagnosis and treatment (Ali *et al*., 2023). Accident-avoidance systems powered by AI within self-driving cars may one day be an effective way to reduce road accidents and protect human life (Khan *et al*., 2022). In manufacturing, AI-driven automation is to some extent transforming production, procurement and logistics processes, increasing productivity and reducing costs (Dwivedi *et al*., 2021a; Richey *et al*., 2023). AI technology has transformed the retail sector via the use of high levels of customer personalisation, behavioural analytics and recommendation engines, to deliver greater levels of consumer satisfaction (Jaheer Mukthar *et al*., 2022). Within finance, the technology is used to analyse market trends and develop insights based on numerous variables and historical data (Hentzen *et al*., 2022). The adoption of AI offers increasing levels of productivity and job satisfaction as relatively mundane and repetitive tasks can be automated. A trial with 5,179 customer support agents within a major software company highlighted that less-experienced workers achieved a 35% rise in productivity and that AI-assisted interactions reduced worker attrition by 8.6% and improved customer satisfaction (NBER, 2023). Organisations that have invested in AI have realised the benefits. Similar outcomes have been reported by Octopus Energy in the UK (Deliotte, 2023).

Generative AI (GenAI) may represent a significant step in the ability of machines to emulate and generate human-like content. The rapid growth of GenAI since the release of OpenAI’s ChatGPT in November 2022, and subsequent launches of Large Language Model (LLM) based products from established “tech” companies such as Meta, Microsoft and Alphabet, has driven a transformation in how we interact with AI and use the technology (Dwivedi *et al*., 2023a; 2023b). Google’s integration of a language model with more than 25 of their existing products, including mail and maps, and the incorporation of image-based prompts in addition to text, highlights how such tools are likely to become useful (Heikkila, 2023). The integration of voice and image capabilities within conversational models increases the capability of interaction with such technology.

The automation and democratisation of AI based content creation, exacerbates the concerns about faked information and associated spread of misinformation (Pawelec, 2022). Models trained on biased data already are recognised to perpetuate societal injustice in the context of crime and finance (Dwivedi *et al*., 2021a). As the penetration of the technology spreads, its governance of AI will require regulation (Wirtz *et al*., 2022; Dwivedi *et al*., 2023a). Though we know already that this is difficult or even impossible to implement it in an international scenario and with the focus on generating revenue.

This article is by contributors to the 4th Royal Society Yusef Hamied Workshop on Artificial Intelligence for India and the UK, held on 24-25 July 2023 in Delhi, India, where scientists from the UK and India discussed the leading research topics that might inspire future research opportunities and in particular collaborations.

The next section lists the individual contributions of six scientists from a variety of technology focussed disciplines. Each has endeavoured to explore a different aspect of the subject and its impact on science, business and society. The Discussion section outlines the key themes, followed by a concluding section.

2. Perspectives

This study aligns with previous opinion based multi-perspective editorials as originally set out in Foerster's (2003) research and numerous subsequent studies that have developed an expert-based perspective on a range of topics including AI and ChatGPT (Budhwar *et al*., 2023; Dwivedi *et al*., 2021a; 2023a; 2023b), metaverse (Dwivedi *et al*., 2023c), impact of digital technologies on climate change (Dwivedi *et al*., 2022), digital and social media marketing (Dwivedi *et al*., 2021b), COVID19 and Information Management (Dwivedi *et al*., 2020), and IS success and failure (Dwivedi *et al*., 2015). We examine the emerging insights on AI and the impact from GenAI and LLMs to develop future perspective and unique insights to the impact from the greater adoption of AI technology. The full list of experts (who participated in the workshop and contributed to this editorial) together with their specific topics and bios are listed in Table 1.

Table 1: Expert contributions who participated in the workshop[2]

Sect #	Contribution Title	About Contributors
2.1	Biomedical Text Summarisation in the Era of Large Language Models	**Sophia Ananiadou** is Professor in Computer Science at The University of Manchester. Her main areas of research are Natural Language Processing with emphasis in Biomedicine. She is the Director of the UK National Centre for Text Mining, Deputy Director of the Institute of Data Science and AI (Manchester), Turing Fellow, ELLIS member, and Distinguished research fellow at the AI research centre (AIST Japan). Co-instigator of the Special Interest Group (SIGBioMed) dedicated to language processing in the biomedical, and clinical domain bringing together researchers in NLP, bioinformatics, and medical informatics
2.2	Evaluating the capabilities of commonsense reasoning in Large	**Anthony G Cohn** is Professor of Automated Reasoning at the University of Leeds, and is also Foundational Models Theme lead at the Alan Turing Institute. He is a Fellow of the Royal Academy of Engineering, and is also a Fellow of AAAI, AISB, EurAI, and AAIA. He is Editor-in-Chief of the

[2] Contributors are listed in alphabetical order according to their surname.

	Language Models	journal Spatial Cognition and Computation. He received the inaugural Test-of-time KR Classic Paper Award in 2020, and the 2021 Herbert A Simon Prize for Advances in Cognitive Systems. He has been given Distinguished Service awards from both IJCAI and AAAI.
2.3	Data-driven materials discovery	**Jacqueline M. Cole** holds the BASF/Royal Academy of Engineering Research Chair in Data-Driven Molecular Engineering of Functional Materials, at the Cavendish Laboratory, University of Cambridge. She is also part-seconded to the UK government through the Science and Technology Facilities Council (STFC). She leads a 20-strong research group in Molecular Engineering at Cambridge, with a focus on developing and applying a range of chemistry-aware AI-based software tools and machine-learning algorithms to solve key materials challenges in the energy sector.
2.4	Data availability and practitioner training for trustworthy AI	**Gareth Conduit** has a track record of applying artificial intelligence to solve real-world problems. The approach, originally developed for materials design, is now being commercialized by startup Intellegens in not only materials design, but also healthcare and drug discovery. Previously, Gareth had research interests in strongly correlated phenomena, in particular proposing spin spiral state in the itinerant ferromagnet that was later observed in CeFePO. Gareth's group is based at the University of Cambridge.
2.5	Explainability in Dialog Systems: Need, Challenges, and Research Directions	**Maunendra Sankar Desarkar** is an Associate Professor at the Computer Science and Engineering Department and also at the Department of Artificial Intelligence at IIT Hyderabad, India. Prior to Joining IIT Hyderabad, he has worked for Samsung Research India Bangalore and Sybase Inc. Dr. Maunendra's main research areas are Natural Language Processing, Information Retrieval, and Machine Learning. He is a part of the Natural Language and Information Processing (NLIP) Research Group at the Department of CSE, IIT Hyderabad. He has several publications in these domains at multiple top tier publication venues.
2.6	Leveraging AI for safer and more efficient transport systems	**Xinwei Wang** is a Lecturer in Systems Engineering at Queen Mary University of London (QMUL), UK. He was a Postdoc at TU Delft, The Netherlands from 2020 to 2022 and at QMUL from 2019 to 2020, respectively. He obtained a PhD degree from Beihang University, China in 2019. Over the years, he has integrated computational intelligence, machine learning and systems engineering for risk assessment, motion planning and decision making in intelligent systems.

2.1 *Biomedical Text Summarisation in the Era of Large Language Models* – ***Sophia Ananiadou***

The huge amount of unstructured biomedical knowledge conveyed in various documents, including scientific literature, electronic health records, clinical notes and clinical trial documentation, presents a significant challenge for researchers who wish to understand and make use of it. Biomedical text summarization (BTS) automatically condenses the content of either single or multiple biomedical documents into concise summaries that capture the most salient information contained within them. These summaries can save researchers a considerable amount of time and effort, since they make it possible to rapidly grasp the main ideas conveyed by long biomedical texts. BTS approaches may be divided into *extractive* and *abstractive* methods. Extractive methods work by selecting key sentences from the original documents and concatenating them into a summary, while in abstractive approaches, summaries consist of *newly generated* sentences, whose content is based on the original documents. Compared to extractive summarization, the task of abstractive summarization is significantly more challenging, since the generation of novel informative sentences involves selecting appropriate words from large vocabulary, as well as syntactic adjustment and paraphrasing. Moreover, generated sentences should be factually consistent with the original text. BTS methods have the potential to be integrated in a variety of real-world applications, including those that aid in the production of systematic reviews for evidence-based medicine, clinical information management systems and decision support systems.

The evolution of pre-trained language models (PLMs) and large language models (LLMs) has revolutionized the field of NLP and has facilitated significant advances in BTS research, leading to the emergence of numerous novel methods, datasets, and evaluation metrics. PLMs are language models that have been pre-trained with large amounts of unlabeled data in a self- supervised manner. The value of PLMs lies in their ability to capture common sense and lexical knowledge that is inherent in the training data. Many BTS methods encode input texts using domain-specific PLMs, such as BioBERT, BlueBERT, and ClinicalBERT. The process of fine- tuning ensures that knowledge captured by the PLMs can be exploited for BTS in an effective manner, such that generated summaries are as informative and coherent as possible. LLMs, which have more parameters than PLMs, are able to achieve higher performance than PLMs, without the need for supervised training. They exhibit a remarkable capacity for both natural language understanding and generation. LLMs also have the ability to carry out in- context learning, i.e., they can perform previously unseen tasks simply by following natural language instructions, either without any specific training data (i.e., zero-shot learning) or by using only a small number of training samples (i.e., few-shot learning). There has recently been a growing interest in exploring the use of LLMs for BTS. These approaches can be broadly divided into three main categories, i.e., data augmentation-based methods, zero-shot-based methods, and domain adaptation-based methods, according to their strategy for leveraging LLMs. Data augmentation-based methods use the generation capabilities of LLMs to automatically augment existing training corpora with additional high-quality training data, which can help to enhance the performance of supervised summarization methods. Zero-shot (and few-shot) methods use natural language instructions to directly prompt LLMs to generate biomedical summaries, without explicit training on the target summarization task. Domain adaptation-based methods focus on continual pre-training or instruction-based fine-tuning of publicly available LLMs using biomedical datasets or task-specific datasets. These approaches aim to allow LLMs to better capture medical knowledge and task-specific information, thus enhancing their effectiveness in carrying out BTS tasks.

Despite their proven effectiveness for BTS, the use of PLMs and LLMs brings a number of challenges. One of these concerns the limitation in the token length of input documents that is imposed by many PLMs and LLMs. This means that long biomedical texts, such scientific papers, can only be encoded by truncating them, which could result in the loss of information that is vital for accurate summarization. However, there are a number of promising paths that could help to address this issue, including techniques that encode the global semantics of documents (Xie, Huang, Saha, & Ananiadou, 2022) and hybrid extractive- abstractive methods that capture the salience of documents (Bishop, Xie, & Ananiadou, 2022).

A further potential issue of using PLMs and LLMs is that although they are able to capture lexical and common-sense knowledge in biomedical texts, they are not aware of which words or entities have particular domain-specific importance, or which types of relations exist between them. However, this type of domain-specific biomedical knowledge, which is encoded in vocabularies, taxonomies, and ontologies such as UMLS, is critical to facilitate a complete understanding of biomedical texts. Indeed, although methods based on PLMs and LLMs can generate summaries that are fluent and largely grammatically correct, their tendency to hallucinate information or contain factual errors is likely to be due, at least in part, to the limited biomedical knowledge that is captured by them. While only a small number of previous studies (e.g., Xie, Bishop, Tiwari, & Ananiadou, 2022; Xie, Tiwari, & Ananiadou, 2023) have investigated the integration of external domain-specific knowledge in BTS methods, these studies have shown that such knowledge can enhance the performance of PLM and LLM- based BTS methods. As such, the development of further knowledge-aware models, which incorporate additional domain-specific knowledge from sources such as UMLS, constitutes a promising path for future investigation.

Although ensuring the factual correctness of automatically generated summaries is a critical requirement for real-world applications of BTS methods, it remains an unresolved challenge, resulting in the danger that generated summaries may contain fabricated facts that do not occur in the original input text. To encourage the development of BTS methods that have a greater emphasis on ensuring factual correctness and coherence of generated summaries, it is important that evaluation metrics move beyond traditional accuracy-based measures, and instead consider a wider range of attributes that contribute towards the overall quality of generated summaries. In this respect, research that focuses on rating factual consistency and readability (Luo, Xie & Ananiadou, 2022; Luo, Xie & Ananiadou, 2023) based on prompt engineering constitutes an important line of investigation. ChatGPT has shown great potential to act as a factual inconsistency evaluator, outperforming state-of-the-art evaluation metrics on widely used datasets. Nevertheless, issues of hallucination, bias and false reasoning need to be further examined. Moreover, the extent to which such automated methods can accurately assess the quality of summaries, in comparison with human evaluation approaches requires further investigation and validation.

A final challenge of employing PLM and LLM-based methods concerns their black-box nature. This lack of transparency means that end-users will often struggle to comprehend the reasons why a model has selected specific words or sentences in the summary generation process. This can be particularly problematic when the model consistently generates errors since it is impossible for users to pinpoint why these inaccuracies occur. Nevertheless, this issue remains largely unexplored in the current research landscape. To better facilitate the construction of applications that are trusted by users, it is crucial that further research is devoted to developing models that are explainable and transparent, particularly in terms of facilitating an understanding of their inner mechanisms and algorithmic functioning.

2.2. Evaluating the capabilities of commonsense reasoning in Large Language Models - ***Anthony Cohn***

Research in *Foundation Models* (FMs) and in particular *large language models* (LLMs) (Zhou *et al*., 2023) has progressed rapidly in recent years (Yang *et al.*, 2023). These are generative models which have been trained on huge datasets and are able to generate text (LLMs) or other kinds of output such as images, video and speech in the case of FMs with other modalities. The best known LLM, certainly by the public at large, is probably ChatGPT (Peters *et al*., 2023) which can produce very fluent and usually coherent text, in many languages, about almost any topic. The performance of LLMs on well-known benchmarks, such as BIG-Bench (Srivastava *et al*., 2022) often reaches "human level", though there are many problems with equating performance on a benchmark with real capabilities on real life data and inputs. As Burnell *et al*. (2023) argue, the cases where a system produces inaccurate responses may be particularly biased towards certain kinds of instance and a full evaluation should document all instances of data and breakdown performance in a more granular way than just a single metric across the whole benchmark as is generally done.

One area which has traditionally been very challenging for AI systems to demonstrate competence in is that of commonsense reasoning. While reasoning about specialist domains is important for many AI applications, the ability to reason reliably and robustly about common sense is also important for many applications, including, for example customer service chatbots. Reasoning about common sense has been a goal of AI since its earliest days (McCarthy, 1959) but it has proved surprisingly hard to endow computers with this capability. Whilst many systems have been built which display expert abilities in fields as varied as medicine (e.g. van Melle, 1978), engineering (Dimitrova *et al.,* 2020), bioinformatics (Tunyasuvunakool, 2021) and even games such Go (Silver *et al*., 2016), there has not yet been a system which has truly displayed the commonsense reasoning abilities of, say, a ten year old child. The question I wish to address here is whether LLMs can already or will be able to display common sense reasoning. The most long-standing effort towards developing common sense in a computer has been the CYC project (Lenat *et al*., 1990) which has taken a symbolic approach and has more than 24M rules and assertions in its ontology and knowledge base and has been used in many applications. However, there has been no comprehensive evaluation of its capabilities and apart from limited subsets (OpenCYC and ResearchCYC (Ramachandran, 2005)) it remains a closed product and does not natively have a natural language interface.

The question arises as to what is common sense? Davis (2023) suggests the following criteria for common sense: (i) is common; (ii) is "largely sensible"; (iii) supports reasoning; (iv) is integrated into other cognitive abilities (language, vision, etc.) - one never observes it directly, only how it is manifested through language, action, etc.; (v) is independent of any modality or task; (vi) has broad scope; (vii) is distinguished from common knowledge, encyclopaedic and expert knowledge; (viii) is concerned with generalities rather than individuals; (ix) is not book learning or explicitly taught in schools; (x) is separate from purely linguistic or purely perceptual interpretation. Davis goes on to present a list of commonsense benchmarks and analyses these as to whether they meet the criteria above (and some other desirable features for benchmarks) and concludes "many of the commonsense benchmarks that have been created do not at all respect these boundaries; most involve substantial amounts of what is clearly common knowledge and many involve rather obscure encyclopaedic knowledge or, more rarely, even expert knowledge". Thus, the oft-reported success of AI systems for being able to reason about common sense because of their performance on such benchmarks is questionable.

A challenge for the future therefore is to determine better ways of evaluating the performance of LLMs, especially in the field of common-sense reasoning, where there are no standard exams or ways of assessing performance (unlike in professional fields). One way of achieving this may be to use a more interactive form of assessment, that has been called *dialectical evaluation* (Cohn & Hernández-Orallo, 2023) – see also Collins *et al*. (2023). In such an evaluation, the assessment of the AI system takes the form of a dialogue, in which subsequent prompts to the LLM depend on the output from the previous one(s). Thus, the extent of the "understanding" of the LLM can be probed and ambiguities in responses resolved and the explanations of responses from an LLM can be analysed and further tested – often an answer may be correct but the proffered explanation faulty, or a presentation of reasoning steps may be perfect but actual answer to a question posed may be incorrect. As the conversation progresses, the previous prompt-response pairs become part of the context for the LLM which provides a richer environment for evaluation than predetermined, standalone, usually multiple-choice questions in benchmark (from which there is always a risk that there has been leakage into the test set (Kocijan, 2022)). Another benefit of the dialectical approach is that variant problems can be easily formulated and evaluated – this is related to what McCarthy (2003) has called *elaboration tolerance* i.e. the ability of a representation to act as a foundation for solving variants of the original challenge problem, though here, since the LLM is a black box and its internal representation is inscrutable, the question is simply whether success on one problem can be replicated in variant or extended problems.

As an example of such a dialectical approach, consider the following conversation with ChatGPT-4 conducted in March 2023, taken from Cohn and Hernández-Orallo (2023). Underlining indicates erroneous responses from the LLM, and italic text indicates that the response is not actually incorrect but has some issues.

> Human: The book couldn't stand upright in the bookcase because it was too small. What does "it" refer to?
>
> ChatGPT-4: In this sentence, "it" refers to the book. The sentence is saying that the book was too small to stand upright in the bookcase.
>
> Human: Why do you say that?
>
> ChatGPT-4: I say that because, based on the context provided in the sentence, the pronoun "it" is used to replace a noun mentioned earlier in the sentence. In this case, the only noun mentioned is "the book." Therefore, "it" logically refers back to "the book," describing its size as being too small to stand upright in the bookcase.
>
> Human: Surely the book wouldn't fit properly only if it was too big, not too small?

> ChatGPT-4: You are correct that if a book is too big, it might not fit properly in a bookcase. However, the original sentence stated that the book was too small, which implies that it might not stand upright due to its size. In this context, the issue might be that the book is too short or too thin to remain upright without additional support. It could lean, tilt, or fall over, instead of standing upright like the other books in the bookcase.
>
> Human: But I didn't talk about falling over, only about fitting.
>
> I apologize for the confusion. Given the context of fitting, if the book is too small, it would still fit in the bookcase, but it might not stand upright as easily or look as visually appealing among the other books. The sentence describes the book as being unable to stand upright, which could mean it doesn't "fit" well in terms of alignment or aesthetics, rather than a physical limitation of space.
>
> Human: in the sentence "The book couldn't stand upright in the bookcase because it was too big", what does the “it" refer to?
>
> ChatGPT-4: In this sentence, "it" refers to the book. The sentence is saying that the book was too big to stand upright in the bookcase.

In Cohn and Hernández-Orallo (2023), the case of common-sense spatial reasoning is investigated across several LLMs and whilst the newer LLMs (in particular ChatGPT4) show better performance in both answer accuracy and explanation correctness, it is clear that spatial reasoning still presents a substantial challenge to LLMs.

A problem that is endemic to LLMs, since they are ultimately just statistical models of language (Bender *et al*., (2021) have called them “stochastic parrots”), is the issue of what has been called “hallucination”, i.e., “inventing facts”, though using the term hallucination may be ascribing more agency to an LLM than is warranted. This problem is also present in commonsense reasoning; for example in a query the author posed recently to Google’s Bard LLM asking it to describe a route based on an extract given to it from a travel guide, it invented road numbers which were not present in the original text, and which were not appropriate to the requested route. A second challenge for such statistical models is whether they can ever truly show the ability to do anything other than rather shallow reasoning, where the reasoning patterns have already been seen, and learned from the training data.

Whilst having many advantages as argued above, dialectical evaluation has the disadvantage that it is very expensive in human effort as it requires a human actively involved “in the loop” during the evaluation. It may be possible to at least partly automate the process, for example by exploiting a virtual world setting in which questions can be posed and answers evaluated in an automated way, though this still presents many challenges to achieve in a robust manner. However, if the purpose of dialectical reasoning is to find the “failure modes” of LLMs, then human-in-the -loop evaluation may be less of an issue, especially since a statistical evaluation is less important. It may also be possible to test sensitivity to linguistic and semantic perturbations in the prompts in a (semi-)automated way. The sensitivity of LLMs to such perturbations has already been noted in a Theory-of-Mind setting (Ullman, 2023). Finally, it is worth noting the predilection of humans to anthropomorphise when interacting with a computer – when presented with fluent and apparently correct responses, there is a natural human tendency to ascribe greater understanding and intelligence to the system than is warranted – see Bundy (2017) for further discussion on this issue in the context of whether smart machines are a threat to humanity or not.

It is also worth pointing out the challenge of evaluating the capabilities of proprietary LLMs, which are only accessible via a web interface or an API. The lack of transparency of both the model and the data it was trained on, make assessing the performance of the LLM particularly challenging. Such LLMs have been termed Language-models-as-a-service (LMaaS) (La Malfa et al., 2023), who discuss why this paradigm presents particular challenges to accessibility, replicability, reliability, and trustworthiness (ARRT) of LMaaS, and who make suggestions for how to ameliorate the situation and also for further research.

In summary, we conclude that whilst commonsense reasoning is something that does not generally pose a challenge to humans, it appears that even the best large language models struggle to consistently reason about common sense. Moreover, testing the capabilities of LLMs in an efficient but thorough manner is not straightforward, but the method of dialectical evaluation has many advantages, though with a non-trivial human effort involved.

2.3. *Data-driven materials discovery* ***- Jacqueline M. Cole***[3]

It was an honour to join a Royal Society delegation to fly to Delhi, India, in July 2023, where we contributed to the Yusef-Hamied Workshop in Artificial Intelligence at the Indian National Science Academy. Participants evenly spanned a UK-India bilateral exchange of knowledge about the topical area of AI. Talks on AI covered the full range of computer science, materials physics, chemical engineering, chemistry, biology, medicine, pharmacology, management, policy and ethics. The goal was to learn from each other and explore mutually beneficial collaborations. On the one hand, such international knowledge exchange will help to collectively solve the plethora of AI challenges that are currently facing the world; on the other hand, it will help to identify challenges that AI can solve for the world.

A key reflection from the workshop was that while AI alone was not a panacea for all our problems, the symbiosis of AI and human efforts could together solve global challenges in a way that the public could trust. The importance of developing methods that meet responsible AI metrics to gain such trust was stressed. For example, a number of talks featured the opportunities and challenges associated with the global adoption of AI tools such as ChatGPT. Benchmarks that assess different types of reasoning in language were discussed, from common sense and spatial reasoning, to how to detect and interpret sarcasm, humour, metaphors and hyperbole in linguistics. Challenges associated with unstructured text were addressed, in terms of how natural-language processing and language models are being used to provide the necessary structure. The need for language models to be adapted for specialised domains of interest was highlighted in biological and materials-science subjects. The use of lightweight transformers for language models was advocated in the field of

[3] **Funding Source:** Jacqueline M. Cole would like to thank the Royal Academy of Engineering for the Research Chair in Data-driven Molecular Engineering of Functional Materials, and the Science and Technology Facilities Council (STFC) for their provision of time to attend this meeting.

biology. Considerations were given to the energy cost of generating large language models, which were relayed in a currency of CO_2 emissions. One the one hand, such energy requirements highlighted the need to create language models that are publicly available. On the other hand, this spirit of open innovation was tempered by requirements in the medical sector to comply with federated learning, i.e., the training of an AI model on multi-source, decentralised, medical data which cannot be shared due to privacy or copyright regulations. This topic aligned with more general debate about the ethics of AI and the desperate need to develop more policy work in this field.

My own research contribution to this workshop demonstrated an AI method that we have developed to auto-generate custom databases for use in a given field of materials research, to suit a desired application. This presentation was built upon the following premise.

Large-scale data-mining workflows are increasingly able to predict successfully new chemicals that possess a targeted functionality. The success of such materials discovery approaches is nonetheless contingent upon having the right data source to mine, and algorithms that suitably encode structure-function relationships as data-mining workflows which progressively short list data toward the prediction of a lead material for experimental validation.

My talk described how to meet these data and algorithmic requirements via a 'design-to-device' approach to data-driven materials discovery (Cole, 2020). The presentation included a description of how to auto-generate large material databases of relevant experimental information from scientific documents, using natural-language processing, language models, computer vision, and machine learning, via our home-grown software toolkits: ChemDataExtractor (Swain & Cole, 2016; Mavracic, Court, Isazawa, Elliott, & Cole, 2021; Isazawa & Cole, 2022; Isazawa & Cole, 2023), ImageDataExtractor (Mukaddem, Beard, Yildirim, & Cole, 2020; Yildirim & Cole, 2021), ChemSchematicResolver (Beard & Cole, 2020) and ReactionDataExtractor (Wilary & Cole, 2021; Wilary & Cole, 2023).

My talk also illustrated how large auto-generated databases of chemical structures and their cognate properties can be mined for materials discovery using custom-built algorithms (Beard, Sivaraman, Vázquez-Mayagoitia, Vishwanath & Cole, 2019; Beard & Cole, 2022; Huang & Cole, 2020). It was also shown how the resulting predictions could be screened using machine-learning algorithms until lead candidate materials appear, which are then experimentally validated. The high-performance photovoltaic device that was afforded experimentally from our predicted lead materials demonstrated the power of this data-driven materials discovery (Cooper *et al.*, 2019).

Our use of language models to generate databases about optical and battery devices was also showcased (Zhao, Huang & Cole, 2023; Huang & Cole, 2022a) as was the ability to create property-specific text-mining tools (Huang & Cole, 2022b).

Having discussed the opportunities of this approach, in terms of these case studies from the energy sector, I also mentioned key challenges that exist within this field. In particular, I advocated the use of small language models (SLMs) that are created from textual corpora that are domain rich in the subject area of interest, where this is practical; rather than using large language models that have been trained on more generic source data. I also recommended the provision of open-source language models and codebases in the spirit of open innovation and to help democratize AI across the world.

AI can be used as a force for good and it could offer transformational opportunities in many aspects of our lives, and its ability to function with low resource needs means that it can be a game changer for developing countries. AI innovations could literally save lives and offer the prospect of saving the planet; from using AI to improve global health, mitigate climate change and reduce poverty. The democratisation of AI will help to solve these challenges, especially the latter issue, if one can find appropriate routes for the education of AI.

The digital revolution is here to stay and it is likewise important to best capture the business prospects of AI for developing countries, since the wealth of a nation can help to offset the above issues to realise positive socio-economic impact. The interplay between creating value for a nation and trying to distribute it equitably across the country, especially in cases where it has a very unequal political power structure, will be a challenge; but not one that we should avoid. The development of regulation and risk policies may help, as might distribution points.

Highly collaborative ways of working are needed to help realise solutions to these complex issues together with a cross-cutting team of the sort that was present at this workshop.

The innate creative and innovative thinking that flowed in and out of the sessions at this workshop was incredible. This editorial is just one of the first of what will hopefully be many collaborative outcomes that will emanate from our time in India.

2.4. *Data availability and practitioner training for trustworthy AI* ***-Gareth Conduit***

The Royal Society Yusuf Hamied Workshop for India and the UK was a highly informative event. I spoke about developments in the application of machine learning to experimental data spanning the materials, chemicals, and pharmaceutical sectors. Two algorithmic developments were highlighted: firstly, the ability to handle inevitably sparse data owing to not every quantity having been measured (Conduit *et al.*, 2017); and secondly the proper estimation of uncertainty due to both experimental statistical errors and extrapolation to new space, and how to account for that uncertainty in robust design (Conduit et al., 2018). These were illustrated by two real-life case studies where artificial intelligence was used to design formulations, whose properties were later experimentally verified.

The talks and discussion during the workshop highlighted two opportunities for future research: the availability of data, and practitioner training & trustworthiness that are discussed further below. These are not only important topics of academic interest, but moreover their development would drive future adoption and impact of artificial intelligence.

2.4.1. Availability of data

During the workshop both developers and users of artificial intelligence tools emphasized that a core requirement for successful model building is a corpus of data to train and validate the model. Over the years organizations have collated and curated data, but jealously guard the data owing to its commercial importance and sometimes sensitive nature. However, it can be mutually beneficial for organizations to share the knowledge as this will, for example, merge complementary information to paint a fuller picture. Two possible approaches for this were discussed:

2.4.1.1. Common ontology and data access: Organizations are sometimes happy to share data publicly, for example academics. However, to make data as useful as possible for artificial intelligence it is essential to bring the data into a single database. The meeting discussed how this requires that data be stored in a common ontology, and be accessible with common calls (Andersen *et al.*, 2021).

2.4.1.2. Federated learning: Where data is necessarily private and cannot be moved outside of the owner's silo, federated learning algorithms allow an individual model to be trained behind the firewall of each data owner, The algorithm can then merge these models to collate the knowledge from private data silos to build a model that learns from all information available, which will then make better predictions and enable better design. There are several promising starts for translation of this approach into industry.

2.4.2. Practitioner training and trustworthy artificial intelligence

The meeting heard about several impactful real-life uses of artificial intelligence. However, despite demonstrable success, a significant barrier to the more widespread adoption of artificial intelligence is the lack of understanding of the approach and concern about the trustworthiness of predictions. Two future research directions were discussed at the meeting:

2.4.2.1. Training: To motivate and enable the adoption of artificial intelligence several complementary approaches could be followed. The publication of exemplar and inspirational real-life use-cases that cover a broad range of disciplines would inspire future users in how they could successfully apply artificial intelligence. Moreover, providing an accompanying clear and demonstrable return on investment on projects would motivate commercial adoption of artificial intelligence. Further adoption of artificial intelligence would be accelerated by expanded Master's courses, particularly those with placements to train future users and spread best practice.

2.4.2.2. Trustworthiness: Many potential users are concerned about how trustworthy predictions from artificial intelligence are. Improving trustworthiness would not only encourage further use of artificial intelligence, but explainability and robustness are essential requirements to pass regulatory hurdles in safety critical applications such as pharmaceuticals and healthcare. The meeting discussed future opportunities to improve trustworthiness as part of ongoing research programs in responsible and trustworthy AI.

2.5. *Explainability in Dialog Systems: Need, Challenges, and Research Directions* **- *Maunendra Sankar Desarkar***

Dialogue systems or Conversational AI agents are becoming increasingly popular over the past few years. These adoptions are mostly backed by promising potentials and commercial values of such systems. With the advancements in LLMs and related research, the quality of the generated responses in dialogue systems has been further enhanced. This is, in turn, increasing the potential of such systems. In this interesting phase, both academia and industry are working towards making such systems more accurate, acceptable, and affordable.

However, there are several concerns regarding the adoption of such dialogue systems. One of these major reasons is the black-box nature of generating the responses. As the responses generated by conversational systems are generally consumed by an end user, the explainability of the generated responses is of utmost importance. Without explainability, the confidence in the generated responses is low. Poor quality responses with a lack of explanation can hit user satisfaction and also hurt business and/or relations. Hence, explainability in dialogue systems requires a special focus - from the perspectives of both modelling and evaluation.

2.5.1. Challenges

Early dialog systems were rule-based. For rule-based systems, it is easy to give a reasoning behind the generated responses. However, the ability to define the rules and having the user-inputs adhere to the rules limits the usability of such systems. Modern dialog systems make use of neural encoder-decoder based models to generate the response - where the utterances can be arbitrary, and the models are still able to handle them due to the powerful neural architectures driving the response generation. However, the black-box nature of the neural models makes it difficult to explain the responses generated (Li *et al*., 2023).

2.5.2. Opportunities

With conversational interfaces gaining popularity, more and more data is being generated for dialog systems research. While some of this data will reside inside the organizational boundaries, some of the data (created for open research or released after obtaining user consent) will be available in the public domain. The availability of such datasets provides an opportunity to look closely into the data and identify scenarios where explanations are missing, incorrect, or inadequate. This can lead to an increased focus on explainable dialog systems research.

2.5.3. Research Agenda

Dialog systems can be broadly divided into two types: Goal-oriented and Non-goal-oriented or chit-chat or open-domain (Luo *et al*., 2019). In Goal-oriented dialog systems, the automated agent tries to help the user achieve a task. For open domain or chit-chat kind of dialog systems, the conversations are typically engaging in nature and need to match a common theme for the discussions. For both types of dialog systems, being able to provide justifications for what has been uttered by the system improves the user's confidence in the system and encourages long-term association. This is useful for both the user as well as the system.

2.5.3.1. Dialog state tracking: For goal-oriented dialog systems, the dialog state contains the mentions of *domain slot values* that are currently under consideration (Dey & Desarkar, 2021). For example, if a user is trying to book a ticket for 3 persons from a specific location X to a specific location Y, these details of number of passengers, source, and destination should be captured in the dialog state. However, if any of these slots (e.g. number of passengers) or slot-values (e.g. number of passengers=3) are not captured (missed), or an incorrect value is assumed, then the entire process may go wrong. Hence, it is important to have components in the system that can provide accurate justification for the predictions made by the system.

2.5.3.2. Knowledge-Grounded Dialogs: In knowledge-grounded dialog systems, based on the current context of the dialog, appropriate knowledge components are selected (Li *et al.*, 2022). These knowledge components can be sentences or paragraphs from a repository. The final response is generated based on the selected knowledge. Having the response grounded on the knowledge can help towards explainability of the generated response. However, in this scenario also, it is important to be able to explain why a specific piece of knowledge segment is selected, as this becomes the basis for the response generation in the subsequent step.

2.5.3.3. Grounding of responses on important context: Although for goal-oriented systems, the dialog states take care of the grounding of the responses, for open-domain, personal conversation, or chit-chat kind of systems, such dialog states are not available. In such situations, it might be better to understand how humans generate responses, and similar strategies can be incorporated in designing the response generation models (Dey *et al*., 2023). It is widely assumed that users keep a contextual summary of the dialog in mind, and also remember some specific utterances or words mentioned in the conversation thus far. Mimicking human behaviour by identifying these summaries, important utterances, and words, and generating the responses conditioned on these aspects can possibly lead to better quality responses.

2.5.3.4. Explaining the dialog policy or act: In Goal-oriented Dialog systems, dialog policy (Rastogi et. al.) determines the next action of the system. For example, it may provide the user with some information that s/he has requested (restaurant in a certain location), or also request some additional inputs or clarifications (cuisine type, distance from an attraction, etc.) to be able to provide a better quality response. It may also offer new intent (booking a taxi, etc). For non-goal-oriented dialog systems, the system may also decide to emphasise upon aspects related to empathy, persuasion, etc. Explainability in such actions will be helpful when the conversation goes wrong or the user is dissatisfied with the responses. The explanations will enable future course corrections or retraining/re-designing of the models if necessary.

2.5.3.5. Explainability in dialog evaluation: Automatically evaluating machine-generated responses is critical and challenging for developing dialog systems. Although there has been tremendous progress in dialogue systems research, the evaluation heavily depends on human judgments (Dey & Desarkar, 2023). The standard word-overlapping-based evaluation metrics, such as BLEU, METEOR, etc. are ineffective for dialogues, as, for the same dialog context, the correct and acceptable response can be given in multiple different ways. So, although a high word-overlap-based score can indicate good performance, a relatively low score for low word-overlap does not necessarily indicate that the response is poor. For promoting dialog systems research and identifying and encouraging good models for further development, the metrics should be designed accordingly. In addition to that, if there are more metrics that can favour explainable methods (Dey & Desarkar, 2023), then the research on explainable dialog generation can be strengthened.

2.6. Leveraging AI for safer and more efficient transport systems - Xinwei Wang

2.6.1. Overview and challenges

AI technologies have been widely applied to various transport systems for safer and more efficient performance (Abduljabbar *et al.*, 2019). A well-known example is automated driving, which heavily relies on AI-enabled processes, including data fusing, sensing, safety measurement and motion planning. This section aims to provide a concise overview of the use of AI for transport systems in terms of safety and efficiency, followed by an exploration of associated challenges and future research opportunities. While we focus on the ground-based transport systems, it is worth noting that other modes of transportation, such as air and space systems, may encounter similar challenges and research avenues.

To improve transport safety, one common approach is to measure the safety by considering interactions between transport participants in a microscopic level (Lefèvre *et al.*, 2014). Traditionally various time- and distance-based safety metrics have been designed to quantify such risk. For instance, Time-To-Collision assumes constant speeds for all vehicles involved and derives a time estimate until collision, and Stopping Distance is determined based on known vehicle speed and maximum deceleration capability. Nevertheless, most time- and distance-based safety metrics adopt a deterministic approach and do not account for the inherent motion uncertainties associated with other road participants. An alternative category of safety metrics, namely probabilistic metrics, offers an avenue to introduce these uncertainties by calculating a collision probability between transport participants. This has served as a base for employing AI-driven trajectory and control input predictors, which anticipate and represent motion uncertainties by utilising deep learning networks, e.g., LSTM and Transformer, resulting in an accurate risk estimation. These networks have been used extensively to augment the effectiveness of probabilistic safety metrics (Wang *et al.*, 2022). From a system-level perspective, there has also been a trend of employing AI for the prediction and management of traffic flows, aimed at mitigating the occurrence of safety-critical scenarios.

While AI has shown potential to address transport safety, it has also been demonstrated an effective tool to improve transport system efficiency (Fadlullah e*t al*., 2017). Firstly, as more transport data is readily available since the digitalisation of transport infrastructure, AI has been implemented for data collection, processing and analysis, resulting in precise time-spatial system metrics distributions, transport pattern classification and performance evaluation outcomes. Secondly, these results can further serve as inputs for transport network management problems, e.g., logistics routing and scheduling, ride-sharing optimisation and truck network platooning coordination. Based on the input data, AI has been employed and combined with other decision-making approaches to train a transport management agent and provide data-driven solutions to various system management problems. Thirdly, AI has also found applications in generating authentic system data and expediting the simulation and validation of transport systems (Feng *et al*., 2023).

As we increasingly rely on AI to enhance transport safety and efficiency, ensuring the robustness of these AI-driven solutions becomes paramount. Two specific questions emerge: **Q1**: To what extent can we depend on the reliability of AI-enabled safety systems? and **Q2**: Is AI capable of enhancing system efficiency across diverse scenarios despite the lack of complete datasets? The answer to Q1 not only affects transport safety but also has profound impact on the overall acceptance and adoption of AI-driven safety approaches in transport systems. As for Q2, we also contemplate the transformative potential of AI in improving transport system efficiency across diverse scenarios, when faced with limited datasets.

Overall, we aim to empower AI to make informed decisions and optimise transport systems when confronted with novel or unforeseen circumstances.

2.6.2. Future research

We respond to the two challenges (Q1 and Q2) by pointing out the following future research on the use of AI in transport systems.

2.6.2.1 Robust and Explainable AI: Future research needs to focus on creating AI models and algorithms that are resilient to unexpected and adversarial conditions. Robust AI will address these issues and perform effectively under various scenarios, from extreme weather conditions to unforeseen traffic disruptions, thereby enhancing safety and efficiency. Besides, AI explainability is a critical factor in the acceptance and adoption of AI-driven technologies in transportation. Future explainable AI aims to make these systems more transparent and understandable to both experts and the public. This involves developing techniques with more interpretable and clearer explanations for the decisions made by AI algorithms. For instance, explainable AI will help passengers and regulators understand why certain routes or decisions are chosen, enhancing transparency and accountability while fostering user confidence.

2.6.2.2. Digital Twins: Digital twins are virtual replicas of physical transport systems, offering real-time insights and analysis capabilities. Future research in this area seeks to advance the fidelity and utility of digital twins. To address fidelity, digital twins will integrate sensor data and AI algorithms, thus provide predictive maintenance, optimise traffic flow, and simulate critical scenarios. It is also essential to explore transfer learning and domain adaptation, which can generalise well from smaller datasets. To address the utility of future digital twins of transport systems, existing transport systems will be revolutionised by allowing for more realistic critical scenarios generation and reproduction and integrating with a proactive-and-reactive decision-making framework.

2.6.2.3. AI Ethics: The issue of AI ethics has been well recognised and discussed (Hagendorff, 2020) in the domain of natural language processing, computer vision and decision support systems. While in transport systems, ethical AI is closely related to ethical dilemmas such as data privacy of transport participants, and routing and scheduling algorithmic fairness among individuals and companies. Ethical AI ensures that decisions made by AI models do not discriminate against certain groups, and that transport data is collected in a distributed manner without violating data regulations.

In summary, AI has made a contribution towards enhancing the safety and efficiency of transportation systems, while robustness and reliability of these AI-driven solutions becomes paramount. By incorporating digital twins and addressing AI robustness, explainability and ethics, future AI-driven transport systems are to be not only technologically advanced and reliable, but also socially trustworthy.

3. Discussion

The impact of widespread adoption of AI has emerged as one of the most significant technological transformations that has the capacity to impact almost every facet of human life and work. The individual perspectives detailed in the previous sections have each offered perspective on many of the key debates and important topics surrounding the use and adoption of AI within an industrial and societal context.

The individual perspectives from the invited contributors to this article have each focussed on specific important areas of AI, each offering valuable perspective on their topic. A number of the contributions referenced the issue of explainability within AI.

The concept of explainability refers to the ability to understand and interpret the decisions and predictions made by machine learning models and AI based systems. This is an active research area where studies have identified the many challenges and impact stemming from poor levels of explainability, and how AI systems can be developed to acceptable levels of transparency to deliver the necessary insight to algorithmic based decision making (Dennehy *et al*., 2023; Dwivedi *et al*., 2021a). Research from McKinsey finds that companies investing and seeking ROI from AI, are more likely to follow best practices that enable explainability, and organisations that engender trust amongst consumers by making AI more transparent and explainable, are more likely to see revenues grow at rates of 10% or more (McKinsey, 2023). The contribution from *Gareth Conduit* references explainability in the context of essential requirements to meet regulatory requirements and its criticality within AI applications in pharmaceuticals and healthcare. These arguments are further developed in the contribution by *Maunendra Sankar Desarkar,* where the discussion articulates how confidence in AI responses are directly impacted by poor levels of explainability, and the need for a special focus on explainability within dialogue systems from both a modelling and evaluation perspective. The contribution calls for a greater focus on explainability in this important area and notes the reliance on humans in the loop within existing research on automatic evaluation techniques (Dey & Desarkar, 2023). The contributions from *Sophia Ananiadou* and *Xinwei Wang* call for robust and explainable AI asserting that transparency and explainability are critical factor in the acceptance and adoption of AI-driven technologies, calling for new techniques for AI algorithms that are more interpretable giving clearer explanations for their decisions.

The principles of trust, safety, and ethics are essential aspects of AI design, development and deployment. Researchers have identified the importance of transparency, reliability, and immediacy as behaviours in developing cognitive trust in AI, and how emotional trust in AI is influenced by anthropomorphic factors (Dwivedi *et al*., 2021a; Glikson & Woolley, 2020). The contributions from *Jacqueline Cole* and *Gareth Conduit* and *Xinwei Wang* all reference trust, highlighting the importance of trustworthiness in the outputs of AI and that this factor is significant barrier to further adoption levels from users. Sectors such as pharmaceuticals must take steps to mitigate the current lack of trust in AI applications, as well as the inadequate regulations needed to protect patient privacy and rights (Murdoch, 2021). In healthcare, transparency related to AI and algorithmic decision making are crucial due to the potential impact on people's lives (Kiseleva *et al*., 2022). The contribution from *Xinwei Wang* discusses the safety implications of AI within transport systems, highlighting the potential for the technology to improve transport system efficiency and enhance the robustness and reliability of automated safety systems, despite the lack of complete datasets (Fadlullah *et al*., 2017; Feng *et al*., 2023). The ethical dimensions of AI and ChatGPT have been widely discussed within the literature (Dwivedi *et al*., 2021a; 2023a; Hagendorff, 2020; Stahl & Eke, 2024). Ethics within an AI context is critical to ensure that the technology serves the best interests of users, and that fairness, transparency, accountability, and privacy are central to the design and deployment of AI systems. The contribution from *Xinwei Wang* highlighted the ethical dilemmas within automated transport systems (Awad *et al*., 2018) and called for AI designers and developers to ensure that decisions made by AI models take account of fairness and discrimination factors. Researchers have referenced these topics in the context of responsible AI, highlighting the role of decision makers in ensuring that AI technologies are trustworthy, safe, and ethical in their deployment (Dignum, 2019).

The significant disruption to numerous business sectors from the development and widespread adoption of LLM's and GenAI based tools and products, has been transformational impacting various aspects of technology, industry and wider society (Dwivedi *et al*., 2023a). Studies have argued that the use of LLMs has a negative impact on areas such as public health. The research by De Angelis *et al*. (2023) highlights the potential for large amounts of generated scientific articles, fake news, and misinformation as a direct consequence of LLM's ability to rapidly generate human-like content without any scientific grounding and demonstrable audit trail. Although LLMs are incredibly powerful, limitations and specific complexities exist when applied to specific domains and industries such as medicine, law, environment, or engineering, where the technology may lack the detailed knowledge and training to provide accurate or specific and nuanced responses to queries (Hadi *et al*., 2023; Ufuk, 2023). The contribution from *Sophia Ananiadou* discusses these complexities in the context of BTS, highlighting that the automatic truncating of long biomedical texts and scientific papers, could result in the loss of vital information and loss of critical accuracy. The topic of AI hallucination is also discussed in this contribution, where BTS outputs may contain fabricated data that is integrated with the scientific text summary. The topic of LLM's is also extensively discussed in the contribution from *Anthony Cohn*, where the discussion focusses on the commonsense reasoning aspect of LLMs, impact of hallucination and the natural tendency for humans to ascribe elevated levels of understanding and intelligence to systems than is justified. Researchers have commented on the impact of LLM adoption in the context of influence and persuasion where no humans are involved in the information generation process. These behaviours have been documented specific to number of policy issues such as an assault weapon bans, carbon taxation, paid parental-leave programs (Bai *et al*., 2023). The extent of humans needed in the loop for AI is an evolving research area within a spectrum of near full automation to integrating multidisciplinary teams in the loop for complex medical processes and procedures (Sezgin, 2023). What is clear is that the level of human intervention in AI needs to be carefully formulated depending on context and threats to human safety and security.

Central to the further development of responsible AI is the further democratisation of the technology to ensure AI is not limited to a select group of academics or experts but widely accessible to empower individuals within developed and developing economies (Ahmed & Wahed, 2020). The democratization of AI entails the design and development of no-code or minimal code, user-friendly AI tools requiring minimal technical expertise. This will effectively open up new areas in medicine, farming, sustainability and small business where AI can be used for repetitive tasks such as data analysis and content generation. The contribution from *Jacqueline Cole* discusses the underlying factors on democratisation of AI calling for greater access to open-source language models and further levels of democratisation of AI to help solve challenges such as improvements in health, mitigating the effects of climate change and to reduce global poverty. The provision of adequate low-cost practitioner training is crucial in ensuring that AI technology receives widespread adoption and is able to fulfil its potential in solving many of the current global challenges. Sectors such as healthcare could be greatly impacted in terms of patient outcomes where decision makers invest in the practitioner training necessary to deliver benefits (Choudhury & Asan, 2022).

Each of the contributions from the invited experts offers a distinct and insightful discussion on core aspects of AI, its significant potential, many complexities and risks to industry and society. Policy makers and decision makers are advised to assess the implications of these discussions and seek to responsibly develop the potential for AI to engender real change at an industrial and societal level.

A number of potential future research directions emerge from this study that centre around the key elements of the contributions and identified challenges stemming from the use and application of AI. These are outlined within Table 2 below.

Table 2: Future Research Directions

Research direction		Contributor
1	Development of models that are explainable and transparent, particularly in terms of facilitating an understanding of their inner mechanisms and algorithmic functioning that do not require human judgments	Sophia Ananiadou Maunendra Sankar Desarkar
2	Research to improve the ability of LLMs to reason, particularly about commonsense, for example by using neuro-symbolic methods.	Anthony Cohn
3	Analysis of improved testing methods for robustly testing LLMs through dialectical evaluation that minimizes human effort and involvement.	Anthony Cohn
4	Research into feasibility of SLMs that are trained in specific domain data that could offer improved levels of specificity and accuracy.	Jacqueline Cole
5	Advocation of greater levels of AI democratisation via open-source language models and codebases.	Jacqueline Cole
6	Research into improving levels of trustworthiness via advances in explainability and robustness – especially in critical areas such as pharmaceuticals and healthcare.	Gareth Conduit Maunendra Sankar Desarkar
7	Development of new databases to store data in a common ontology that would be widely accessible with common API's to further AI research.	Gareth Conduit
8	New approaches to federated learning algorithms that will allow an individual model to be trained behind the firewall of each data owner.	Gareth Conduit
9	Research into goal-oriented and non-goal-oriented dialogue strategies that include accurate justification for the predictions made by the system.	Maunendra Sankar Desarkar
10	Research into dialogue systems behaviour that can better explain why a specific piece of knowledge segment is selected.	Maunendra Sankar Desarkar
11	Further research into use of AI in transport systems to ensure that decisions made by AI models are ethical and do not discriminate against certain groups, and that data is collected in a distributed manner without violating data regulations.	Xinwei Wang
12	Research to develop techniques with more interpretable and clearer explanations for the decisions made by AI algorithms in transport systems.	Xinwei Wang
13	Development of new digital twin simulations to revolutionise existing transport systems by allowing for more realistic critical	Xinwei Wang

	scenario generation and reproduction within an integrated proactive-and-reactive decision-making framework.	

4. Conclusions

This study has been developed from the scientific discussions and the leading research topics from the 4th Royal Society Yusef Hamied Workshop on Artificial Intelligence for India and the UK. The individual contributions have been distilled within this article where we have discussed the many emerging challenges and considerations stemming from greater diffusion and adoption of AI technology. The contributions highlight the impact from AI within industry and society and illustrate how the use of this technology could further the democratisation of AI but also potentially create mistrust and fear of the technology where AI designers and developers fail to deliver adequate levels of transparency and explainability of AI algorithmic decision making. What is clear is that the use of AI can deliver significant change at a global level but it requires the ingenuity and ideas of humans to direct this to the biggest problems that will deliver the most benefit to mankind and the planet.

Acknowledgements: *Authors of this editorial wish to extend their sincere gratitude to the Royal Society and the Indian National Science Academy for convening the 4th Yusuf Hamied Workshop, which brought together a diverse group of thought leaders and experts. The interactions and discussions that took place during the workshop played an instrumental role in shaping the ideas presented in this editorial. While the Royal Society and the Indian National Science Academy did not participate in the development or publication of the editorial, we deeply appreciate their commitment to fostering dialogue and collaboration within the scientific community*.

References

Abduljabbar, R., Dia, H., Liyanage, S., & Bagloee, S. A. (2019). Applications of Artificial Intelligence in Transport: An Overview. *Sustainability*, 11(1), Article 1. https://doi.org/10.3390/su11010189

Ahmed, N., & Wahed, M. (2020). The De-democratization of AI: Deep learning and the compute divide in artificial intelligence research. *arXiv preprint arXiv:2010.15581*.

Ali, O., Abdelbaki, W., Shrestha, A., Elbasi, E., Alryalat, M. A. A., & Dwivedi, Y. K. (2023). A systematic literature review of artificial intelligence in the healthcare sector: Benefits, challenges, methodologies, and functionalities. *Journal of Innovation & Knowledge*, *8*(1), 100333.

Andersen, C.W., Armiento, R., Blokhin, E., Conduit, G.J. , Dwaraknath, S., Evans, M.L. , … & Yang, X. (2021). OPTIMADE: an API for exchanging materials data. *Nature Scientific Data* 8, 217.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... & Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59-64.

Bai, H., Voelkel, J., Eichstaedt, J., & Willer, R. (2023). Artificial intelligence can persuade humans on political issues.

Beard, E. J. & Cole, J. M. (2020). ChemSchematicResolver: a toolkit to decode 2D chemical diagrams with labels and R-groups into annotated chemical named entities. *Journal of Chemical Information and Modeling, 60*(4), 2059-2072. www.chemschematicresolver.org.

Beard, E. J. & Cole, J. M. (2022). Perovskite- and dye-sensitized solar-cell device databases auto-generated using ChemDataExtractor. *Scientific Data, 9*(1), 329.

Beard, E. J., Sivaraman, G., Vázquez-Mayagoitia, Á., Vishwanath, V., Cole, J. M. (2019). Comparative dataset of experimental and computational attributes of UV/vis absorption spectra. *Scientific Data, 6*(1), 307.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).

Bishop, J., Xie, Q., & Ananiadou, S. (2022). GenCompareSum: a hybrid unsupervised summarization method using salience. Proceedings of the 21st Workshop on Biomedical Language Processing, 220–240, Association for Computational Linguistics.

Budhwar, P., Chowdhury, S., Wood, G., Aguinis, H., Bamber, G. J., Beltran, J. R., ... & Varma, A. (2023). Human resource management in the age of generative artificial intelligence: Perspectives and research directions on ChatGPT. *Human Resource Management Journal*, 33(3), 606-659.

Bundy, A. (2017). Smart machines are not a threat to humanity. *Communications of the ACM*, 60(2), 40-42.

Burnell, R., Schellaert, W., Burden, J., Ullman, T. D., Martinez-Plumed, F., Tenenbaum, J. B., ... & Hernandez-Orallo, J. (2023). Rethink reporting of evaluation results in AI. *Science*, 380(6641), 136-138.

Choudhury, A., & Asan, O. (2022). Impact of accountability, training, and human factors on the use of artificial intelligence in healthcare: Exploring the perceptions of healthcare practitioners in the US. *Human Factors in Healthcare*, *2*, 100021.

Cohn, A. G., & Hernandez-Orallo, J. (2023). Dialectical language model evaluation: An initial appraisal of the commonsense spatial reasoning abilities of LLMs. *arXiv preprint arXiv*:2304.11164.

Cole, J. M. (2020). A design-to-device pipeline for data-driven materials discovery. *Accounts of Chemical Research*, *53*(3), 599–610.

Collins, K. M., Jiang, A. Q., Frieder, S., Wong, L., Zilka, M., Bhatt, U., ... & Jamnik, M. (2023). Evaluating Language Models for Mathematics through Interactions. *arXiv preprint arXiv*:2306.01694.

Collins, K. M., Jiang, A. Q., Frieder, S., Wong, L., Zilka, M., Bhatt, U., ... & Jamnik, M. (2023). Evaluating Language Models for Mathematics through Interactions. *arXiv preprint arXiv*:2306.01694.

Conduit, B.D., Jones, N.G., Stone, H.J., & Conduit, G.J. (2017). Design of a nickel-base superalloy using a neural network. Materials & Design 131, 358.

Conduit, B.D., Jones, N.G., Stone, H.J., & Conduit, G.J. (2018). Probabilistic design of a molybdenum-base alloy using a neural network. Scripta Materialia 146, 82.Cooper, C. B., Beard, E. J., Vázquez-Mayagoitia, Á., Stan, L., Stenning, G. B. G., Nye, D. W., Vigil, J. A., Tomar, T., Jia, J., Bodedla, G. B., Chen, S., Gallego, L., Franco, S., Carella, A., Justin Thomas, K. R., Xue, S., Zhu, X. & Cole, J. M. (2019). Design-to-device approach affords panchromatic co-sensitized solar cells. *Advanced Energy Materials*, *9*(5), 1802820.

Davis, E. (2023). Benchmarks for automated commonsense reasoning: A survey. *arXiv preprint arXiv*:2302.04752.

De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G. P., Ferragina, P., Tozzi, A. E., & Rizzo, C. (2023). ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Frontiers in Public Health*, *11*, 1166120.

Deliotte (2023). Artificial Intelligence – Jobs crash, productivity boom? *Accessed on 26th September 2023*. https://www.deloitteacademy.co.uk/node/4566

Dennehy, D., Griva, A., Pouloudi, N., Dwivedi, Y. K., Mäntymäki, M., & Pappas, I. O. (2023). Artificial intelligence (AI) and information systems: perspectives to responsible AI. *Information Systems Frontiers*, *25*(1), 1-7.

Dey, S., & Desarkar, M. S. (2021, July). Hi-DST: A hierarchical approach for scalable and extensible dialogue state tracking. In Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue (pp. 218-227).

Dey, S., & Desarkar, M. S. Dial-M: A Masking-based Framework for Dialogue Evaluation. 24th Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2023).

Dey, S., Desarkar, M. S., Ekbal A., & Srijith P. K. DialoGen: Generalized Long-Range Context Representation for Dialogue Systems. 37th Pacific Asia Conference on Language, Information and Computation (PACLIC 37).

Dignum, V. (2019). *Responsible artificial intelligence: how to develop and use AI in a responsible way* (Vol. 2156). Cham: Springer.

Dimitrova, V., Mehmood, M. O., Thakker, D., Sage-Vallier, B., Valdes, J., & Cohn, A. G. (2020). An ontological approach for pathology assessment and diagnosis of tunnels. *Engineering Applications of Artificial Intelligence*, 90, 103450.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., ... & Wright, R. (2023a). "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*, *71*, 102642.

Dwivedi, Y.K., Pandey, N., Currie, W., & Micu, A. (2023b). Leveraging ChatGPT and other generative artificial intelligence (AI)-based applications in the hospitality and tourism industry: practices, challenges and research agenda. *International Journal of Contemporary Hospitality Management*, doi: https://doi.org/10.1108/IJCHM-05-2023-0686

Dwivedi, Y. K., Hughes, L., Wang, Y., Alalwan, A. A., Ahn, S. J., Balakrishnan, J., ... & Wirtz, J. (2023c). Metaverse marketing: How the metaverse will shape the future of consumer research and practice. *Psychology & Marketing*, 40(4), 750-776.

Dwivedi, Y. K., Hughes, L., Kar, A. K., Baabdullah, A. M., Grover, P., Abbas, R., ... & Wade, M. (2022). Climate change and COP26: Are digital technologies and information management part of the problem or the solution? An editorial reflection and call to action. *International Journal of Information Management*, 63, 102456.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., ... & Williams, M. D. (2021a). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, *57*, 101994.

Dwivedi, Y. K., Ismagilova, E., Hughes, D. L., Carlson, J., Filieri, R., Jacobson, J., ... & Wang, Y. (2021b). Setting the future of digital and social media marketing research: Perspectives and research propositions. *International Journal of Information Management*, 59, 102168.

Dwivedi, Y. K., Hughes, D. L., Coombs, C., Constantiou, I., Duan, Y., Edwards, J. S., ... & Upadhyay, N. (2020). Impact of COVID-19 pandemic on information management research and practice: Transforming education, work and life. *International journal of information management*, 55, 102211.

Dwivedi, Y. K., Wastell, D., Laumer, S., Henriksen, H. Z., Myers, M. D., Bunker, D., ... & Srivastava, S. C. (2015). Research on information systems failures and successes: Status update and future directions. *Information Systems Frontiers*, 17, 143-157.

Fadlullah, Z. Md., Tang, F., Mao, B., Kato, N., Akashi, O., Inoue, T., & Mizutani, K. (2017). State-of-the-Art Deep Learning: Evolving Machine Intelligence Toward Tomorrow's Intelligent Network

Traffic Control Systems. *IEEE Communications Surveys & Tutorials*, 19(4), 2432–2455. https://doi.org/10.1109/COMST.2017.2707140

Feng, S., Sun, H., Yan, X., Zhu, H., Zou, Z., Shen, S., & Liu, H. X. (2023). Dense reinforcement learning for safety validation of autonomous vehicles. *Nature*, 615(7953), 620–627. https://doi.org/10.1038/s41586-023-05732-2

Foerster, H. V. (2003). On self-organizing systems and their environments. In Understanding Understanding (pp. 1-19). Springer, New York, NY.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627-660.

Goldman Sachs (2023). Generative AI could raise global GDP by 7%. Accessed on 26th September 2023. https://www.goldmansachs.com/intelligence/pages/generative-ai-could-raise-global-gdp-by-7-percent.html

Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., ... & Mirjalili, S. (2023). A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage. *TechRxiv*.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, *30(1)*, 99-120.

Heikkila., M (2023). MIT Technology Review: Google is throwing generative AI at everything. Accessed on 11th May 2023. https://www.technologyreview.com/2023/05/10/1072880/google-is-throwing-generative-ai-at-everything/

Hentzen, J. K., Hoffmann, A., Dolan, R., & Pala, E. (2022). Artificial intelligence in customer-facing financial services: a systematic literature review and agenda for future research. *International Journal of Bank Marketing*, *40*(6), 1299-1336.

Huang S. & Cole, J. M. (2020). A database of battery materials auto-generated using ChemDataExtractor. *Scientific Data, 7*(1), 260.

Huang, S. & Cole, J. M. (2022a). BatteryBERT: A pretrained language model for battery database enhancement. *Journal of Chemical Information and Modeling*, *62*(24), 6365-6377.

Huang, S. & Cole, J. M. (2022b). BatteryDataExtractor: battery-aware text-mining software embedded with BERT models. *Chemical Science*, 13(39), 11487-11495.

Isazawa, T., Cole, J. M. (2022). Single model for organic and inorganic chemical named entity recognition in ChemDataExtractor. *Journal of Chemical Information and Modeling, 62*(5), 1207-1213. www.chemdataextractor2.org (version 2.1).

Isazawa, T., Cole, J. M. (2023). Automated Construction of a Photocatalysis Dataset for Water-Splitting Applications. *Scientific Data,* 10, 651. https://doi.org/10.1038/s41597-023-02511-6. www.chemdataextractor2.org (version 2.2).

Jaheer Mukthar, K. P., Sivasubramanian, K., Ramirez Asis, E. H., & Guerra-Munoz, M. E. (2022). Redesigning and Reinvention of Retail Industry Through Artificial Intelligence (AI). In *Future of Organizations and Work After the 4th Industrial Revolution: The Role of Artificial Intelligence, Big Data, Automation, and Robotics* (pp. 41-56). Cham: Springer International Publishing.

Khan, S., Adnan, A., & Iqbal, N. (2022, July). Applications of Artificial Intelligence in Transportation. In *2022 International Conference on Electrical, Computer and Energy Technologies (ICECET)* (pp. 1-6). IEEE.

Kiseleva, A., Kotzinos, D., & De Hert, P. (2022). Transparency of AI in healthcare as a multilayered system of accountabilities: between legal requirements and technical limitations. *Frontiers in Artificial Intelligence*, *5*, 879603.

Kocijan, V., Davis, E., Lukasiewicz, T., Marcus, G., & Morgenstern, L. (2023). The defeat of the Winograd schema challenge. *Artificial Intelligence*, 103971, doi: https://doi.org/10.1016/j.artint.2023.103971

La Malfa, E., Petrov, A., Frieder, S., Weinhuber, C., Burnell, R., Cohn, A. G., ... & Wooldridge, M. (2023). The ARRT of Language-Models-as-a-Service: Overview of a New Paradigm and its Challenges. arXiv preprint arXiv:2309.16573.

Lefèvre, S., Vasquez, D., & Laugier, C. (2014). A survey on motion prediction and risk assessment for intelligent vehicles. *ROBOMECH Journal*, 1(1), 1–14.

Lenat, D. B., Guha, R. V., Pittman, K., Pratt, D., & Shepherd, M. (1990). Cyc: toward programs with common sense. *Communications of the ACM*, 33(8), 30-49.

Li, S., Sun, C., Xu, Z., Tiwari, P., Liu, B., Gupta, D., ... & Wang, M. (2023). Toward Explainable Dialogue System Using Two-stage Response Generation. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(3), 1-18.

Li, Y., Peng, B., Shen, Y., Mao, Y., Liden, L., Yu, Z., & Gao, J. (2022, July). Knowledge-Grounded Dialogue Generation with a Unified Knowledge Representation. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 206-218).

Luo, L., Huang, W., Zeng, Q., Nie, Z., & Sun, X. (2019, July). Learning personalized end-to-end goal-oriented dialog. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 6794-6801).

Luo, Z., Xie, Q., & Ananiadou, S. (2022). Readability Controllable Biomedical Document Summarization. Findings of the Association for Computational Linguistics: EMNLP 2022, 4667–4680.

Luo, Z., Xie, Q., & Ananiadou, S. (2023). ChatGPT as a Factual Inconsistency Evaluator for Text Summarization, https://arxiv.org/abs/2303.15621

Mavracic, J., Court, C. J., Isazawa, T., Elliott, S. R. & Cole, J. M. (2021). ChemDataExtractor 2.0: Autopopulated ontologies for materials science. *Journal of Chemical Information and Modeling, 61*(9), 4280–4289; www.chemdataextractor2.org (version 2.0)

McCarthy, J. (1959). Programs with common sense. In Mechanisation of Thought Processes. *In Proceedings of a Symposium held at the National Physical Laboratory* on 24th, 26th and 27th November 1958. London: H. M. Stationery Office. Vol. 1, pp. 75–84.

McCarthy, J. (2003). *Elaboration Tolerance*. Stanford University

McKinsey (2023). Why Business Leaders Need Explainable AI and How to Deliver It. Accessed on 1st October 2023. https://www.mckinsey.com/capabilities/quantumblack/our-insights/why-businesses-need-explainable-ai-and-how-to-deliver-it

Mukaddem, K., Beard, E. J., Yildirim, B. & Cole, J. M. (2020). ImageDataExtractor: a tool to extract and quantify data from microscopy images. *Journal of Chemical Information and Modeling, 60*(5), 2492-2509. www.imagedataextractor.org (version 1.0).

Murdoch, B. (2021). Privacy and artificial intelligence: challenges for protecting health information in a new era. *BMC Medical Ethics*, *22*(1), 1-5.). Privacy and artificial intelligence: challenges for protecting health information in a new era. *BMC Medical Ethics*, *22*(1), 1-5.

NBER (2023). Generative AI at Work. *Accessed on 26th September 2023*. https://www.nber.org/papers/w31161

OpenAI (2023). ChatGPT can now see, hear and speak. *Accessed on 26th September 2023*. https://openai.com/blog/chatgpt-can-now-see-hear-and-speak

Pawelec, M. (2022). Deepfakes and democracy (theory): how synthetic audio-visual media for disinformation and hate speech threaten core democratic functions. Digital society, 1(2), 19.

Peters, M. A., Jackson, L., Papastephanou, M., Jandrić, P., Lazaroiu, G., Evers, C. W., ... & Fuller, S. (2023). AI and the future of humanity: ChatGPT-4, philosophy and education–Critical responses. *Educational Philosophy and Theory*, 1-35, DOI: 10.1080/00131857.2023.2213437

Ramachandran, D., Reagan, P., & Goolsbey, K. (2005, July). First-orderized researchcyc: Expressivity and efficiency in a common-sense ontology. In *AAAI workshop on contexts and ontologies: theory, practice and applications* (pp. 33-40), (Pittsburgh, PA), AAAI-05.

Rastogi, A., Zang, X., Sunkara, S., Gupta, R., & Khaitan, P. (2020). Schema-guided dialogue state tracking task at DSTC8. arXiv preprint arXiv:2002.01359.

Richey Jr, R. G., Chowdhury, S., Davis□Sramek, B., Giannakis, M., & Dwivedi, Y. K. (2023). Artificial intelligence in logistics and supply chain management: A primer and roadmap for research. *Journal of Business Logistics*, doi: https://doi.org/10.1111/jbl.12364

Sezgin, E. (2023). Artificial intelligence in healthcare: Complementing, not replacing, doctors and healthcare providers. *Digital Health*, *9*, 20552076231186520.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587), 484-489.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., ... & Wang, G. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv*:2206.04615.

Stahl, B. C., & Eke, D. (2024). The ethics of ChatGPT–Exploring the ethical issues of an emerging technology. *International Journal of Information Management*, 74, 102700, doi: https://doi.org/10.1016/j.ijinfomgt.2023.102700

Swain, M. C. & Cole, J. M. (2016). ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature. *Journal of Chemical Information and Modeling*, *56*(10), 1894–1904 (2016); www.chemdataextractor.org

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., ... & Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873), 590-596.

Ufuk, F. (2023). The role and limitations of large language models such as ChatGPT in clinical settings and medical journalism. *Radiology*, *307*(3), e230276.

Ullman, T. (2023). Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv*:2302.08399.

Van Melle, W. (1978). MYCIN: a knowledge-based consultation program for infectious disease diagnosis. *International Journal of Man-Machine Studies*, 10(3), 313-322.

Wang, X., Alonso-Mora, J., & Wang, M. (2022). Probabilistic Risk Metric for Highway Driving Leveraging Multi-Modal Trajectory Predictions. *IEEE Transactions on Intelligent Transportation Systems*, 23(10), 19399–19412. https://doi.org/10.1109/TITS.2022.3164469

Wilary, D. M. & Cole, J. M. (2021). ReactionDataExtractor: a tool for automated extraction of information from chemical reaction schemes. *Journal of Chemical Information and Modeling, 61*(10), 4962-4974. www.reactiondataextractor.org (version 1.0)

Wilary, D. M. & Cole, J. M. (2023). ReactionDataExtractor 2.0: A deep learning approach for data extraction from chemical reaction schemes. *Journal of Chemical Information and Modeling,* https://doi.org/10.1021/acs.jcim.3c00422 [http://www.reactiondataextractor.org/ (version 2.0)]

Wirtz, B. W., Weyerer, J. C., & Kehl, I. (2022). Governance of artificial intelligence: A risk and guideline-based integrative framework. *Government Information Quarterly*, *39*(4), 101685.

Xie, Q., Bishop, J., Tiwari, P., & Ananiadou, S. (2022). Pre-trained language models with domain knowledge for biomedical extractive summarization, Knowledge-Based Systems,Vol. 252,109460, https://doi.org/10.1016/j.knosys.2022.109460.

Xie, Q., Huang, J., Saha, T. & Ananiadou, S. (2022). GRETEL: Graph Contrastive Topic Enhanced Language Model for Long Document Extractive Summarization. Proceedings of the 29th International Conference on Computational Linguistics, 6259–6269

Xie, Q., Tiwari, P., & Ananiadou. S. (2023) Knowledge-enhanced Graph Topic Transformer for Explainable Biomedical Text Summarization, IEEE Journal of Biomedical and Health Informatics, doi: 10.1109/JBHI.2023.3308064.

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., ... & Hu, X. (2023). Harnessing the power of llms in practice: A survey on chatgpt and beyond. *arXiv preprint arXiv*:2304.13712..

Yildirim, B. & Cole, J. M. (2021). Bayesian particle instance segmentation for electron microscopy image quantification. *Journal of Chemical Information and Modeling, 61*(3), 1136–1149. www.imagedataextractor.org (version 2).

Zhao, J., Huang, S. & Cole, J. M. (2023). OpticalBERT and OpticalTable-SQA: text- and table-based language models for the optical-materials domain. *Journal of Chemical Information and Modeling, 63*(7), 1961-1981.

Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., ... & Sun, L. (2023). A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv*:2302.09419.