Concerns about the role of artificial intelligence in journalism, and media manipulation.

Simon Mahony
College of Education for the Future, Beijing Normal University at Zhuhai,
Department of Information Studies, University College London (UCL),
s.mahony@ucl.ac.uk
ORCID: 0000-0001-9811-9381

Qing Chen (corresponding author)
Beijing Normal University
School of International Chinese Language Education
qingchen@mail.bnu.edu.cn
ORCID: 0000-0003-0275-0713

Abstract

Artificial Intelligence is a term used frequently in academic and other writing, but do we have a clear understanding of what it means? This article starts from first principles, taking a dialectic approach, to raise questions rather than give prescriptive answers. It unpacks some specific examples of the use of AI in journalism and automated approaches to news reporting. The manipulation of media has become commonplace and of greater interest as information itself can be used as an effective weapon to sow confusion and disruption, socially as well as politically. AI depends on the training data and modelling, but the sampling and engineering is done by humans with all the potential for bias, whether intentional or not. Biased datasets and the potential for uncertainty are constant dangers; we need to understand both the data and the processes that go into the AI-driven results, and always be prepared to question everything.

Keywords:
Artificial Intelligence, Bias, Communication, Journalism, Media, Social Change.

Introduction

Artificial intelligence (AI) has become a ubiquitous but problematic term used in many different contexts, often without any clear idea of exactly what this term represents (Lewis, 2019). This article asks if we understand what we mean by AI, and whether we should have concerns over the use of this term or indeed of AI itself in specific contexts. With a background in humanities and social science rather than engineering, the authors go back to first principles and that is Socratic Dialectic, to ask questions to arrive at a better understanding, and to test our assumptions. What do we understand by AI and how does it impact on aspects of journalism, media manipulation, and bias, separate but interlinked issues? Our perspective comes from an interest in journalism, information and communication studies, with a particular focus on media and its societal impact; AI has significant implications for journalism in all these areas.

Starting from first principles we need definitions to see how this term is used. The Oxford English Dictionary (OED) Online lists a single entry for artificial intelligence: 'The capacity of computers or other machines to exhibit or simulate intelligent behaviour; the field of study concerned with this. Abbreviated AI.' (OED s.v.: artificial intelligence). An initial question would be what do we understand by intelligence in this context; whether we can imagine the intelligence of a machine to be of the same order of that of a human or, for example, a dog or cat, as far as understanding or intellect are manifest. Further, how might that be simulated and what would we understand by that? Indeed, would the simulation of understanding even be possible and, if it were, would that be the same as real understanding? (Harnard, 1989). Perhaps, it is a different order of intelligence that we might consider here, one that does not engage with consciousness or emotion, one lacking in sentience and cognitive ability (Lavelle, 2020), as well as empathy; one that does not attempt to mimic human behaviour (Scriven, 1953). In its 2020 White Paper, the European Commission puts the emphasis on the technologies used: ‘Simply put, AI is a collection of technologies that combine data, algorithms and computing power. Advances in computing and the increased availability of data are therefore key drivers of the current upsurge of AI’ (European Commission, 2020: 2). Interestingly, the word *trust* is included in the sub-title (- *A European approach to excellence and trust*) which points to concerns that will be picked up later. Moreover, the recently enacted European Parliament’s Artificial Intelligence Act (March 2024), defines AI thus:

> ‘AI system’ means a machine-based system designed to operate with varying levels of autonomy, that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. (European Parliament, 2024. 3(1))

The definitions are extensive and differ according to the users’ perspective.

Although Alan Turing did not coin the term AI, his seminal paper Computing Machinery and Intelligence (1950) introduced the Imitation Game (now popularly known as the Turing Test). He established the concept of a ‘learning machine’, acknowledging that this idea might be ‘paradoxical to some’ as the learning process itself would cause changes to the rules that had been programmed with the teacher consequently often being ‘largely ignorant of what is going on inside the programme’ (Turing, 1950). This prompted much discussion on the fundamentals of what has become known as AI (Mays, 1952; Scriven, 1953), including looking back to Descartes and his Language Tests and Action Tests (Gunderson, 1964; Erion, 2001). In addition, challenges were also set for Turing’s ideas such as by Searle with his Chinese Room Argument for a simulation to counter the Turing Test (Searle 1980). Later ‘opinions on the validity and, especially, the value of the Turing Test as a real guide for research vary widely', particularly if considering intelligence in a philosophical sense (French, 2000: 116). Nevertheless, Turing is considered to be 'father of a lot of modern theory about what computers and computation are' (Haugeland, 1989), and so of computer science rather than AI, although some still consider the Turing Test 'as a benchmark to identify intelligence of an artificial system' (Haenlein, 2019: 3).

The OED gives the first occurrence of the word(s) recorded in the extant English language, which for AI is a paper published by McCarthy et al. (1955), *A proposal for the Dartmouth summer research project on artificial intelligence*. The original proposal is archived and available at Dartford College and Stanford University (McCarthy et al, 1955) with reprinted extracts published in *AI Magazine* (McCarthy et al, 2006). In their words, 'The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.' (McCarthy et al, 1955: 2). They use the term *learning* from the start, as did Turing (1950), and point towards the possibility of 'self-improvement' being an indicator of a 'truly intelligent machine' although importantly that 'the difference between creative thinking and unimaginative competent thinking lies in the injection of some randomness' (McCarthy et al, 1955: 2). Self-improvement and randomness are in their view essential components of successful AI, although it is not clear how that randomness could be 'precisely described'. Turing also acknowledged the wisdom of including 'a random element in a learning machine [as] a random element is rather useful when we are searching for a solution' (Turing, 1950). There clearly needs to be some flexibility, rather than a rigid set of rules, but always within necessary limits: 'Intelligent behaviour presumably consists in a departure from the completely disciplined behaviour involved in computation, but a rather slight one, which does not give rise to random behaviour, or to pointless repetitive loops' (Turing, 1950).

Early influential work in the field of AI was from Newell and Simon 'in the late 1950s through to the early 1990s [with their] crucial contribution to the field, namely symbolic AI', where problems can be represented by human-readable representations (or symbol systems) (Augusto, 2021: 29). Their initial focus on 'problem solving and general intelligence' shifted to 'the nature of representation and knowledge' using symbols (Augusto, 2021: 29-30), although this does not appear to match the conditions for intelligence which have self-improvement and randomness as essential components. Nevertheless, the field has moved on towards machine-learning and now we might consider AI to be an umbrella term that encompasses a wide spectrum:

> we can regard AI as a subfield of the academic discipline of computer science—just as algebra is a subfield of mathematics. Inside AI, there are other subfields: machine learning, expert systems, and natural language processing, to name a few. However, machine learning is the field that is most popular at this cultural moment. When people say "AI" in a business context, generally they are referring to machine learning. Again, machine learning is a subfield of AI, and, like its

> parent, its name misleadingly suggests that sentience exists inside the computer. […] The AI we have today is merely complex and beautiful mathematics. (Broussard et al, 2019: 677)

Mathematics with some randomness then, but, nevertheless, it seems that in general terms within AI, human logic is reversed and that complex (hard) tasks, the ones that are more difficult for humans with structured steps and rules to follow, are easier to programme than the mundane tasks that require less human thought (Minsky, 1986; Moravec, 1988). In what has become known as Moravec's Paradox (Agrawal, 2010), it is the perceptive and intuitive tasks, often achieved unconsciously, that humans find easy and that are the most difficult for AI, whereas logic-based ones are much easier to programme (Aberšek, 2021). Put simply, the computer is better at following rules and counting things than humans and hence these aspects are easier to programme into an AI system.

The term AI has been with us for several decades and appears to have been adopted as the lingua-franca term in other languages too. It has become a field of study, finding its way into academic research and taught university programmes; within the (UK) academic sphere there seems to be a current focus on machine-learning, language processing, education, and knowledge representation.[1] A brief literature search indicates the flexibility and diverse range of academic research that makes use of AI as a topic as well as those utilising it as part of their toolkit (Belle, 2023). Indeed, the term most widely used is simply AI, with the acronym seemingly divorced from the words it represents. Just as so-called Big Data was the ubiquitous solution to all our problems a decade ago, now AI, particularly in combination with Big Data, seems to have taken on a mysticism of its own.

Method

Within the information field, knowledge representation is an important topic; not only how we represent knowledge but also how we understand it mediated by the mode of its presentation. Using linked Open Data and common open standards such as RDF (Resource Description Framework), web-based technologies enable the sharing of discrete data sets via automated and interoperable systems. The data can link freely, but to make it interoperable depends on how the information is represented (hence the need for AI), the data modelling, and use of ontologies. Ontologies, not in the philosophical sense concerned with the nature of being, but rather as part of systems constructing frameworks for the organisation of information, such as the examples of their use in journalism discussed below.

The methodological approach taken here is a dialectic one where questions are raised throughout to examine our understanding of the topics and terms used. The questions we wish to ask are separate but interlinked issues: should we have concerns over the use of AI in specific contexts; how does AI impact on aspects of journalism; how might we mitigate known issues around the manipulation of media, the weaponisation of information, and bias (conscious and unconscious)? Our starting point is journalism and the confidence (or lack of) that we have in the news that is presented to us and we question the validity of some typical uses of AI in journalism. This is followed by examples of AI in media manipulation; problems concerning training models; moves towards regulation and governance; concluding thoughts and final questions finish up this article. Examples and relevant literature are introduced throughout.

Fake News

Within journalism the fabrication of news stories is nothing new and has a long history. It has, however, received more attention recently in both journalism and academic research, particularly following the US Presidential election of 2016 and the UK exit from the European Union. There is significant published research accounting for the phenomenon of AI-generated disinformation and importantly how and why people accept it as truth (Pennycook and Rand, 2021; Waisbord, 2018). Turning again to the OED for a definition, we find:

> fake news n. *originally U.S.* news that conveys or incorporates false, fabricated, or deliberately misleading information, or that is characterized as or accused of doing so. (OED s.v.: fake news, n.)

And further that:

> The term was widely popularized during and after the 2016 U.S. presidential election campaign, and since then has been used in two main ways: to refer to inaccurate stories circulated on social media and the internet, esp. ones which serve a particular political or ideological purpose; or to seek to discredit media reports regarded as partisan or untrustworthy.

Then it goes on to say:

> Some earlier evidence may not represent a fixed collocation, although the practice of 'faking' news stories was much discussed in the late 19th and early 20th centuries (see fake v.2,7a).

The earliest recorded occurrence in the OED Online is the *Milwaukee (Wisconsin) Daily Jrnl*. 7 Feb 1890 with the quote: 'That mine story is one of the greatest pieces of fake news that has been sprung on the country for a long time.'

The Library of Congress Image Catalogue holds an image titled *The fin de siècle newspaper proprietor* (The end of the century newspaper proprietor), attributed to Frederick Opper and dated 1894; this depicts a wealthy newspaper proprietor with cash overflowing from a cabinet marked 'PROFITS', surrounded by images of reporters hastening to deposit their stories, some clearly entitled 'Fake News' (Library of Congress, n.d.). The image makes clear that fake news brings advantage, and in this case a lucrative financial one, but also that both reporters and proprietor are fully aware that the news is fake; it is clearly marked as such. Hence fake news was a recognised part of journalism at that time; nowadays AI plays a significant part in the targeting of fake news stories to those that might be influenced by them.

Things have moved on since then with ‘online platforms, particularly social media […] becoming the main sources of news for a growing number of people’ (Edison et al., 2018: 138). Hence, the dissemination of false or misleading information has become significantly more active in the digital age, with people now tending to dismiss ideas that are contrary to their own as fake news while seeking confirmation bias and ideas that confirm their own beliefs (Gwebu et al., 2022). Consequently, the advent of AI has widened the possibilities for the generation of disinformation and the manipulation of opinion.

Moreover, we have a significant history of propaganda and politically aligned bias purporting to be news predating the spread of social media. This has been demonstrated by Paul Lazarsfeld and colleagues with research on voter choice resulting from ‘selective exposure’ during the 1940 US Presidential campaign and the influence of information from various sources, including the media of that time (Berelson et al., 1948). This activity has taken on new forms and been significantly magnified by the advent of AI, and the ubiquity of social media. In this environment people seem to accept and give credence, consciously or unconsciously, to information that reinforces their views, no matter how distorted or inaccurate, while dismissing content with which they do not agree as fake news. This has a polarising effect and reduces the common ground on which reasoned debate, based on objective facts, can take place (UK Government, 2019: Summary: 5). Fake news, nevertheless, has a range of meanings and to an extent requires audience acceptance otherwise it remains a work of fiction rather having any illusion of truthfulness (Edson et al., 2018). Simply banning publications that are known to be intentionally false is arguably in conflict with freedom of expression and the censorship of the media. In some circumstances, however, government intervention would be necessary to uphold its responsibility to protect national security, public safety, prevent crime and so on (Equality and Human Rights Commission, 2021). What is needed in addition is an informed citizenship educated in critical information literacy (Brisola and Doyle, 2019) to help people have the necessary tools to evaluate and, if needed, challenge the information that is fed to them. This lack of competency with critical examination of information is a global issue with the United Nations General Assembly encouraging:

> all Member States to develop and implement policies, action plans and strategies related to the promotion of media and information literacy, and to increase awareness, capacity for prevention and resilience to disinformation and misinformation, as appropriate. (UN A/RES/75/267, 2021)

Automated journalism

The above is still reported journalism but using AI to manipulate and target specific recipients. There are other forms of journalism that rely almost entirely on AI. This is where articles are generated by computer programmes (AI algorithms) with tasks automated to detect and extract information to produce news content. This is often associated with 'data journalism' and the presentation of tables and charts extracted automatically and often pulled together from disparate datasets, a combination of AI and Big Data. This could mitigate laborious time-consuming tasks and free up journalists to spend their time more productively.

One driver for facilitating these approaches is the development of specific ontologies such as the BBC Storyline Ontology, created in collaboration and released under a Creative Commons Licence.

> **The News Storyline Ontology** is a generic model for describing and organising the stories news organisations tell. The ontology is intended to be flexible to support any given news or media publisher's approach to handling news stories. At the heart of the ontology, is the concept of **Storyline**. As a nuance of the English language the word 'story' has multiple meanings. In news organisations, a story can be an individual piece of content, such as an article or news report. It can also be the editorial view on events occurring in the world. (BBC Storyline Ontology, n.d.)

A philosophical approach to 'ontology' would consider the characterisation of the fundamental nature of existence. 'The science or study of being; that branch of metaphysics concerned with the nature or essence of being or existence.' (OED s.v.: ontology). What can we understand from this about this approach to journalism? Within information studies, we would consider an ontology as creating a structural framework for organising information; hence the Storyline Ontology is released as RDF, the standard model for data exchange on the web. Nevertheless, for AI in the information field, it is necessary to take an epistemological approach and consider the nature of knowledge and of the data that is used. The ontologies allow us to describe the nature of the relationships between the different segments of data/information by allowing classes and properties to be assigned.

From the online description:

> Storyline components can be indisputable real world events, or other storylines (chapters, sub-plots, updates, news developments etc). Storylines can be associated with Topics in some knowledge domain (eg people, places, organisations). (BBC Storyline Ontology, n.d.)

This raises important issues of concern; are any events 'indisputable'? This is questionable as arguably all events are subject to interpretation and potentially open to observational bias, whether unconscious or otherwise.

Another issue here is the use of Large Language Models (LLMs) for automated journalism, using generative AI. The implications regarding the data used in AI applications are addressed later, but LLMs being used for generating text for journalism multiplies the risks of automated articles with their consequent impact on employment and on the potential for editors’ liability. It is also important to foreground Intellectual Property questions linked with their use when the data content is opaque and not open to verification; generative AI models ‘and LLMs in particular, [although they] exhibit high performance across a broad spectrum of tasks […] their unpredictable outputs raise concerns about the lawfulness and accuracy of the generated content.’ (Novelle et al, 2024).

Many other examples could be included here. The Reuters Institute for the Study of Journalism, at the University of Oxford, 'is dedicated to exploring the future of journalism worldwide through debate, engagement, and research.' (Reuters Institute). Published research on their website describes automated journalism in news agencies across Europe. They point to the widespread use in Europe and the USA of machine-generated content, with thousands of stories per month on mainly 'sport and finance' produced with the help of automated algorithms. Acknowledged in their findings is that 'machine written stories lack in-depth and critical examination of the presented facts' (Fanta, 2017). As scholars, we would ask questions and want to verify evidence to come to a reasoned conclusion, but AI algorithms ‘cannot ask questions, determine causality, form opinions’ (Wölker and Powell, 2021: 87). Trust is a significant issue that is at risk, particularly if the publishers do not make it clear by some statement or other indication that these stories are produced algorithmically. Although automated news stories are mostly limited to events in which structured data is commonly available, such as finance and sports (hence the use of ontologies), additional stories are now produced extensively by algorithms for the world's main news agencies: Associated Press, *Agence France Presse*, Reuters and United Press International (Graefe, 2016).

The overall areas of concern fall into several categories: the accuracy of the information, which in turn relies of the accuracy of the data input; can this be verified? Verification is the foundation of scientific (or any scholarly) method. The balance and objectivity of the content that is presented as the news item; are the sources used credible? Is the data presented in a fair way or reliant on subjectivity? Will the point of view expressed serve to cause fragmentation of public opinion? Importantly, would full coverage require a global network for cooperation between newspapers to be developed? Is there any potential for bias, from the data collection, human interpretation, or any intrinsic bias of the algorithms? Technology is never neutral and unconscious bias may be introduced either in the way the algorithms are engineered or in the way that they are implemented (Slate, 2012). We also need to question whether algorithmic journalism would ever report on institutionalised corruption, endemic racism and/or sexism, or government scandals, which would require tenacious investigative journalism instead.

Media manipulation

Modern technology has now enabled what we might consider to be strategic and operational cyberwarfare – the Fifth Domain of conflict (land, sea, air, and space being the previous four domains) (Kirk, 2019). This takes many forms, not just at a technical level with hacking and 'distributed denial of service' attacks but is often journalism targeted to sow confusion and to introduce uncertainty into our decision making, to undermine our trust in our institutions (Bets and Stevens, 2011). This is often characterised as 'Information Warfare', with a range of goals and which has existed as a term in published research for more than two decades (Denning, 1999). The West has a long tradition dating back to the Cold War. More recently we find this definition from NATO which notes the connection 'with the Russian-Ukrainian conflict and the annexation of Crimea by Russia in 2014':

> Information warfare is an operation conducted in order to gain an information advantage over the opponent. It consists in controlling one's own information space, protecting access to one's own information, while acquiring and using the opponent's information, destroying their information systems and disrupting the information flow. Information warfare is not a new phenomenon, yet it contains innovative elements as the effect of technological development, which results in information being disseminated faster and on a larger scale. (NATO, 2014)

This type of conflict has diverse mechanisms and ways of being implemented, low barriers to entry and, above all, 'plausible deniability' as it is often impossible to identify the protagonists. It is a hybrid form of warfare with wide-ranging implications from state interference to Intellectual Property theft, to support strategic industries, to organised crime (Hoffman, 20007). This activity can be personal to an individual or national towards a state; the latter can be viewed as either state interference or state defence, depending on the objective view of the news reporting. There have been many examples of both: the Clinton email hack, deemed to be an Advanced Persistent Threat that occurred prior to the US presidential election in 2016 followed by the elected president's statements regarding NATO and the withdrawal of US support (Bump, 2018). In these cases, journalistic information itself was allegedly used as a weapon enabled by the internet and fuelled by social media.

Such attacks cause us to cast doubts on our knowledge and understanding of a situation; they feed a crisis of confidence in ourselves and in so doing cause disruption (Bets and Stevens, 2011). They raise an awareness of our vulnerability, and we are indeed vulnerable, primarily from our dependence on electronic media and communication. Moreover, we are often unable to identify the protagonists who have multiple and sophisticated mechanisms for concealment. Our infrastructures are fragile which also leaves them vulnerable to attack, but it is the disruption of public confidence and that vulnerability which can have the most profound effect. Attributed to Francis Bacon (1561-1626), English philosopher and statesman, 'Knowledge itself is power' or more correctly '*scientia potentia est*' by Thomas Hobbes (1651), acknowledges that those who have the knowledge (information) also hold the power; information itself has a disruptive power.

Research conducted at the University of Oxford, Computational Propaganda Research Project, identifies and catalogues significant findings concerning major organisations behind government sanctioned social media manipulation, targeting both the domestic and foreign public, much of this being delivered by AI algorithms and bots.

> Social media […] is the primary medium over which young people, around the world, develop their political identities and consume news. However, social media platforms—like Facebook and Twitter—have also become tools for social control. Many governments now spend significant resources and employ large numbers of people to generate content, direct opinion and engage with both foreign and domestic audiences. (Bradshaw and Howard, 2017: 4)

These private organisations, along with big tech companies, have the ability to use their processing of information to influence people or to become censors of the information that reaches us themselves, rather than the State.

Social Media

Media manipulation links directly with social media which has become ubiquitous in our everyday lives; mechanisms for keeping in touch with friends and family, they also act as ready-to-consume digests of news and current events. These aspects are interlinked as users discover and share news and information with the people that they interact with (friends, family, colleagues, etc) as a way of making intersections with others as well as with other parts of the world. This latter aspect has made it the perfect mechanism as a propaganda medium and tool for social control; like it or not, it influences the way we understand the world around us and consequently what we think. When disseminating journalistic information, the word choices, tone of language, and choice of themes can affect the way people perceive the topic. Even if the news media is not successful in telling people what to think, it often has success in telling people what to think about (Cohen, 2015).

Facebook which like many other platforms was designed for entertainment is now, allegedly, used for manipulating public opinion, and that ‘social media platforms [have] emerged as a critical threat to public life’ (Bradshaw and Howard, 2017: 3). When we join our 'friends' on any social media platform they most likely hold the same opinions as we do ourselves. The downside is that we then create our own ‘echo chamber’ of like-minded persons who doubtless also share our own biases and unconscious prejudices, and who, however unwillingly or unwittingly, will tend to reinforce our own worldview (Del Vicario et al., 2017). Hence, this becomes a place where opinion, political leaning, or belief about a topic get reinforced due to repeated interactions with peers or sources with the same attitudes (Cinelli et al., 2021). The AI algorithms reinforce this impression by filtering and delivering content that they think that we would like – more of the same thoughts and ideas as our own. They feed us what we would like to hear, and in extreme cases this can result in politically biased material being targeted to the marginal and undecided voters. Nevertheless, this is part of the business model for social media platforms with the effect of polarisation within our societies (Brown, 2021), concerns which recent proposals by the European Commission seek to address: Digital Services Act[2] and Digital Markets Act[3].

Allegations have been made about politically targeted material being delivered in the runup to both the US presidential election and the UK Brexit referendum in 2016. Much of this is speculation but we need to consider possible motives for such intervention and why this might have been done. Would an isolationist policy in the USA be an advantage to NATO or weaken it and who would benefit from this change? This is not a political science article and so this will not be developed further but, nevertheless, and importantly the question is asked. The same with the Brexit campaign; would the UK leaving strengthen or weaken the European Union and would anyone benefit from this?

Cambridge Analytica is the company that appears to have helped Trump win the 2016 election by harvesting personal information from users' Facebook accounts 'to target them with personalised political advertisements' (Cadwalladr and Graham-Harrison, 2018; New York Times, 2018). Similarly, Facebook appeared to use ‘[s]elective exposure and confirmation bias,’ during the Brexit debate to ‘elicit the formation of polarized groups’ (Del Vicario et al., 2017: 1). In 2018 the UK Information Commissioner's Office (ICO) fined Facebook the maximum possible at the time, 500,000 GBP, 'for serious breaches of data protection law' and 'failing to protect users' personal information' (ICO, 2018).

> We fined Facebook because it allowed applications and application developers to harvest the personal information of its customers who had not given their informed consent […] and then Facebook failed to keep the information safe. […] Facebook broke data protection law […]; it is about the release of users’ profile information without their knowledge and consent. (UK Government, 2019: 21)

Cambridge Analytica, which closed in 2018 following publication of the data scandal, was a UK political consulting firm and a subsidiary of SLC Group (Strategic Communication Laboratories), a contractor for the USA and UK military with strong links to the UK Conservative Party.[4]

The harvested data was allegedly used for targeted advertising using AI algorithms to target the marginal and undecided in the 2016 USA presidential election and the 2016 UK Brexit referendum and, hence, to influence the result. This:

> compounded fears that the [AI] algorithms that determined what people see on the [Facebook] platform were amplifying fake news and hate speech, and that Russian hackers had weaponized them to try to sway the election in Trump's favour. (Hao, 2021)

This was a new generation of AI machine learning algorithms trained on Facebook's datasets, rather than traditional ones hardcoded by engineers. There is concern even within the field itself of the ability to control these algorithms once they are released (Hao, 2021). When training data sets and our data models, we often seek what is generally referred to by machine learning colleagues as the 'ground truth', a term used to refer to data that is somehow known to be real or true (Pickles, 1995). Nevertheless, we need ways to verify this 'ground truth' from both quantitative and qualitative evaluation; the former needs a balance between precision and recall while the latter relies on human verification. Simply using quantitative methods is insufficient for verification as it does not allow the possibility of falsification which is essential for establishing this concept (Popper, 2002).

Training data models

Much depends on the training data used for the AI algorithms. How was the data sampled, what are its limitations, and when collected was it organised to target any specific population demographic? At each stage there is potential for bias to enter the calculation whether that is conscious or unconscious, with the latter often based on culture and upbringing, from which we can never escape. Technology is never neutral; it involves programmers and datasets, with human intervention in both (State, 2012).

In the construction of training data and the engineering of algorithms for AI applications, there is no certainty that the systems are designed, developed, and implemented by experts and programmers who are representative of the diversity of the people who will be most affected by such systems. Without this, there is no assurance of accuracy in the data modelling and hence the AI-delivered predictions. There are many reported examples of flawed models with unpredictable results and some from the UK follow.

Uber Eats, the food delivery sub-section of the taxi company, 'uses Microsoft face-matching software to verify the identity of its couriers when they submit pictures of their own faces' (Kersley, 2021). Problems occurred for their BAME (black, Asian, and minority ethnic) workers attempting to use the automated authentication system. A question to consider here is whether any facial recognition software or identification technology can perform equally across different ethnicities. How have the AI algorithms been engineered and how diverse were the training datasets? In this example the Uber drivers would be uploading 'selfies' taken on their mobile phones which would not be comparable with studio type images held on the company database. What is needed is a multistakeholder assessment of the AI systems and any deployment by employers should be respectful with regards to their employees' ethnicity and the potential for discrimination.

Another example is where the London Metropolitan Police tested facial recognition technology in real-time to monitor crowds at sporting events. Researchers there questioned ethical and privacy concerns regarding this mass application as well as the results themselves (Castelvecchi, 2020). The full report points to an overall lack of guidance on the use of this technology as well as issues that would potentially conflict with human rights law such that its use 'may be held unlawful if challenged before the courts' (Fussey and Murray, 2019: 5). Other studies have shown that machine learning can discriminate based on race and gender; it seems that software responds better to a fairer skin colour and that many training datasets contain significant demographic bias (Buolamwini and Gebru, 2018). If the datasets used for training of the algorithms are biased, then so will the outcomes of any testing and so deliver flawed results. As above, the technology is never neutral as it has been engineered and developed by a person (often male) who carries unconscious cultural and other biases as part of their makeup; 'AI is no different.' (Broussard et al, 2019: 678).

An important issue here, and a concern noted in the European Commission's White Paper, is that of trust. A lack of trust, they claim, 'is a main factor holding back a broader uptake of AI' which prompts their moving towards developing a clear regulatory framework (European Commission, 2020: 9). Trust is often reduced regarding decisions (such as those above) made by AI applications because of a lack of verifiability (Samek and Müller, 2019). Hence the need for what has become known as *explainable AI*; 'in part motivated by the need to maintain trust between the human user and AI' (Jacovi et al, 2021: 624). Whether this trust can be achieved seems uncertain, particularly as the risks to the use of AI in different scenarios are as yet unclear and not fully understood. It maybe that a healthy mistrust is indeed the best approach.

Regulation and governance

As with all aspects of innovative technology, the speed of development exceeds the ability of governments to provide adequate governance. Concerns have been raised within the industry itself regarding the societal impact of the use of AI (Hao, 2021), as well as acknowledgement of potential risks that accompany the many possibilities, particularly resulting from the drive for profitable and lucrative applications (Wirtz et al, 2020).

There have been a significant number of recent government initiatives to draft regulatory frameworks for AI with publications from the UK Department for Science, Innovation & Technology, *AI regulation: a pro-innovation approach* (Gov.uk, 2023) and the European Union, *EU AI Act: first regulation on artificial intelligence* (News European Parliament, 2023). In the USA, Sam Altman, CEO of OpenAI, supported the need for government intervention and regulation of the AI industry (Kang, 2023a), as it appears that lawmakers there are 'far behind Europe' (Kang, 2023b).

The UK document emphasises trust so that the benefits of AI can outweigh the risks, pointing to potential causes of concern that may be in the public perception such as to 'damage our physical and mental health, infringe on the privacy of individuals, and undermine human rights'; 'Public trust in AI will be undermined unless the risks, and wider concerns about the potential for bias and discrimination, are addressed' (Gov.uk, 2023: 4-5). The EU Act seeks to ban what they describe as 'unacceptable risk AI systems' that include:

> Cognitive behavioural manipulation of people or specific vulnerable groups: for example voice-activated toys that encourage dangerous behaviour in children. Social scoring: classifying people based on behaviour, socio-economic status or personal characteristics. Real-time and remote biometric identification systems, such as facial recognition. (News European Parliament, 2023)

In addition, generative AI would need to be fully transparent. The newly enacted European Parliament's Artificial Intelligence Act (2024) emphasises 'the need to build trust' and that it is 'vital' for the regulatory framework 'to be developed in accordance with the Union values'; AI should be human-centric and 'serve as a tool for people, with the ultimate aim of increasing human wellbeing' (6).

We see governments introducing governance to address similar and related concerns around safeguards and building trust in the minds of the public. All seem to have similar concerns. We would question whether regulation alone would be sufficient; regarding AI platforms, there is an absolute need not only to have regulation but overall to have effective and controlled regulation.

Conclusion

Returning to the dialectic method, what do we understand by intelligence in the context of journalism; how might we mitigate issues of media manipulation, the weaponisation of information, and the potential for bias? Awareness and understanding are the first steps. When presented with journalism or articles through our media channels, we need to be clear about the evidence used to underpin claims made and the data used in their support, along with how any sampling was done; this needs to be open and transparent. Then we can have an informed idea about the quality of the data used and hence have a degree of confidence and trust, not only in the data that we are presented with but in the training data that was used to engineer the algorithms. This must be fully documented and open to transparency to avoid the 'black box' scenario where we can only see the inputs and outputs with no understanding or critique of the inner workings, the algorithms and hence how the outputs were generated. Trust is paramount.

Education is also fundamental to this process, and we must always be prepared to question every stage: to question the automated facial recognition, the automatically generated news reports, or the authenticity and reliability of information that comes to us through our social media feeds, or generative AI. This is often covered in the first sessions of library and/or information science programmes under the rubric of data literacy: evaluating data for authenticity and reliability, looking for bias and balance in the information that is presented to us. In many humanities disciplines students are trained to collect and critically evaluate material from a variety of sources and arrive at their own conclusions based on evidence. This brings us back to methodological principles and the scientific method where transparency with fully documented research and experimentation is needed to allow reproducibility and hence the verification of results; this is the foundation on which scholarly practice is built. ‘Science should be “show me”, not “trust me”’ (Stark, 2018), and clarity is always needed.

There are additional ethical concerns, touched on above, that need caution and testing; those are outside the scope of this article but nevertheless are of concern for wider research in the field of AI in journalism. With regards to regulation and governance, many initiatives are in progress globally with moves to engender trust and confidence, but we shall have to wait and see how these develop and whether they are effective and controlled with any substantial impact on the concerns mentioned above. Overall, it is education that leads to understanding that is vital; we need to understand the data being presented and the processes being used, and we should always be prepared to question anything and everything. We should particularly be critical of the data that is being used and look for the verification of the evidence being presented to us; we should always make sure that we ask the right questions. The machines should do what they do best, following rules and counting, and people should do what they do best, exercising perception, intuition, sentience, and empathy. There is plenty of room for both.

References

Aberšek B (2021) The paradox between truth and lies. *Problems of Education in the 21st Century*, 79(6): 834-837.

Agrawal K (2010) To study the phenomenon of the Moravec's paradox. *arXiv preprint* arXiv:1012.3148. Available at https://arxiv.org/abs/1012.3148.

Augusto LM (2021) From symbols to knowledge systems: A. Newell and HA Simon's contribution to symbolic. *AI Journal of Knowledge Structures & Systems* 2(1) 29-62.

BBC Storyline Ontology. Available at: https://www.bbc.co.uk/ontologies/storyline-ontology.

Belle V (2023) Knowledge representation and acquisition for ethical AI: challenges and opportunities. *Ethics In Information Technology* 25(22).

Berelson B, Lazarsfeld PF and Gaudet H (1948) *The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign*. Germany: Columbia University Press.

Bets DJ and Stevens T (2011) *Cyberspace and the State: Towards a Strategy for Cyber-power*. Routledge.

Bradshaw S and Howard P (2017) Troops, Trolls and Troublemakers: A Global Inventory of Organized Social Media Manipulation. *In Computational Propaganda Research Project*: 1–37. Oxford Internet Institute.

Brisola AC and Doyle A (2019) Critical information literacy as a path to resist "fake news": Understanding disinformation as the root problem. *Open Information Science*, 3(1): 274-286.

Broussard M, Diakopoulos N, Guzman AL, Abebe R, Dupagne M and Chuan C H (2019). Artificial intelligence and journalism. *Journalism & mass communication quarterly*, 96(3): 673-695.

Brown S (2021) The case for new social media business models. MIT Management. Available at: https://mitsloan.mit.edu/ideas-made-to-matter/case-new-social-media-business-models.

Bump P (2018) Timeline: How Russian agents allegedly hacked the DNC and Clinton's campaign. *The Washington Post*. Available at: https://www.washingtonpost.com/news/politics/wp/2018/07/13/timeline-how-russian-agents-allegedly-hacked-the-dnc-and-clintons-campaign

Buolamwini J and Gebru T (2018) Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, in *Proceedings of Machine Learning Research* 81: 77-91.

Cadwalladr C and Graham-Harrison E (2018) Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian*. Available at: https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election.

Castelvecchi D (2020) Is facial recognition too biased to be let loose? *Nature* 587: 347-349.

Cinelli M, Morales G, Galeazzi A and Starnini M (2021) The echo chamber effect on social media, *Proceedings of the National Academy of Sciences*, 118(9), e2023301118.

Cohen BC (2015) *Press and foreign poli*cy (Vol. 2321). Princeton university press.

Del Vicario M, Zollo F, Caldarelli G, Scala A and Quattrociocchi W (2017) Mapping social dynamics on Facebook: The Brexit debate. *Social Networks*, 50: 6-16.

Denning D (1999) *Information Warfare and Security.* Addison-Wesley.

Equality and Human Rights Commission (2021) Article 10: Freedom of expression. Available at: https://www.equalityhumanrights.com/human-rights/human-rights-act/article-10-freedom-expression.

Erion GJ (2001) The Cartesian Test for Automatism1. *Minds and Machines*, 11: 29-39.

European Commission (2020) White Paper: On artificial intelligence - A European approach to excellence and trust. Available at: https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.

European Parliament (2024) Artificial Intelligence Act. P9_TA(2024)0138. Available at: https://www.europarl.europa.eu/RegData/seance_pleniere/textes_adoptes/definitif/2024/03-13/0138/P9_TA(2024)0138_EN.pdf.

Fanta A (2017) Putting Europe's Robots on the Map: Automated journalism in news agencies. *Reuters Institute Fellowship Paper University of Oxford*. Available at: https://reutersinstitute.politics.ox.ac.uk/our-research/putting-europes-robots-map-automated-journalism-news-agencies.

French RM (2000) The Turing Test: the first 50 years. *Trends in cognitive sciences*, *4*(3); 115-122.

Fussey P and Murray D (2019) Independent report on the London Metropolitan Police Service's trial of live facial recognition technology. Available at: https://repository.essex.ac.uk/24946/1/London-Met-Police-Trial-of-Facial-Recognition-Tech-Report-2.pdf.

Graefe A (2016) Guide to Automated Journalism. *Tow Center for Digital Journalism*, Columbia University.
Gov.uk (2023) AI regulation: a pro-innovation approach. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachme

nt_data/file/1176103/a-pro-innovation-approach-to-ai-regulation-amended-web-ready.pdf (accessed 02 October 2023).

Gunderson K (1964) Descartes, La Mettrie, language, and machines. *Philosophy*, 39(149): 193-222.

Gwebu K, Wang J and Zifla E (2022) Can warnings curb the spread of fake news? The interplay between warning, trust and confirmation bias. *Behaviour and Information Technology*, 41(16): 3552-3573.

Haenlein M and Kaplan A (2019) A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California management review*, *61*(4): 5-14.

Hao K (2021) MIT Technology Review. How Facebook got addicted to spreading misinformation. Available at: https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation.

Harnad S (1989) Minds, machines and Searle. *Journal of Experimental & Theoretical Artificial Intelligence*, 1(1), pp.5-25.

Haugeland J (1989) *Artificial intelligence: The very idea*. MIT press.

Hobbes T (1651) *Leviathan, Revised Edition*. Martinich and Battiste eds (2010). Broadview Press.

Hoffman F (2007) *Conflict in the 21stCentury: The Rise of Hybrid Wars*. Potomac Institute.

ICO (2018) Investigation into the use of data analytics in political campaigns A report to Parliament. Available at: https://ico.org.uk/media/action-weve-taken/2260271/investigation-into-the-use-of-data-analytics-in-political-campaigns-final-20181105.pdf.

Jacovi A, Marasović A, Miller T and Goldberg Y (2021) Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 624-635).

Kang C (2023a) OpenAI’s Sam Altman Urges A.I. Regulation in Senate Hearing. *The New York Times,* 16 May.

Kang C (2023b) In U.S., Regulating A.I. Is in Its 'Early Days' *The New York Times,* 21 July.

Kersley A (2021) Couriers say Uber's 'racist' facial identification tech got them fired. *WIRED*. Available at: https://www.wired.co.uk/article/uber-eats-couriers-facial-recognition.

Kirk AD (2019) Artificial Intelligence and the Fifth Domain. *Air Force Law Review*. 80: 183-236.

Lavelle S (2020) The machine with a human face: From artificial intelligence to artificial sentience. In *Advanced Information Systems Engineering Workshops: CAiSE 2020 International Workshops*, 32: 63-75.

Lewis SC (2019) Artificial intelligence and journalism. *Journalism & mass communication quarterly*, 96(3), 673-695.

Library of Congress. Available at: https://lccn.loc.gov/2012648704 (accessed 05 June 2023).

Mays W (1952) Can machines think?. *Philosophy*, *27*(101): 48-162.

McCarthy J, Minsky M L, Rochester N, and Shannon, C E (1955) *A proposal for the Dartmouth summer research project on artificial intelligence.* Available at http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf.

McCarthy J, Minsky M L, Rochester N, and Shannon, C E (2006) A proposal for the Dartmouth summer research project on artificial intelligence, august 31, 1955. *AI magazine*, 27(4): 12-14.

Minsky M (1986) *The Society of Mind*. Simon &amp; Schuster. Inc.

Moravec H (1988) *Mind children: The future of robot and human intelligence*. Harvard University Press.

NATO (2014) MEDIA – (DIS)INFORMATION – SECURITY, Information Warfare – NATO. Available at: https://www.nato.int/nato_static_fl2014/assets/pdf/2020/5/pdf/2005-deepportal4-information-warfare.pdf.

News European Parliament (2023) EU AI Act: first regulation on artificial intelligence. Available at:

https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

New York Times (2018) Cambridge Analytica and Facebook: The Scandal and the Fallout So Far. Available at: https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html.

Novelli C, Casolari F, Hacker P, Spedicato G and Floridi L (2024) Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity. *arXiv preprint*. Available at: https://doi.org/10.48550/arXiv.2401.07348.

OED OED s.v. "artificial intelligence, n." OED Online. March 2013. Oxford University Press. Available at: https://www.oed.com/view/Entry/271625.

OED s.v. "fake, n.2 and adj. Compounds C1" OED Online. Available at: https://www.oed.com/view/Entry/67776.

OED s.v. "ontology, n.". OED Online. June 2021. Oxford University Press. Available at: https://www.oed.com/view/Entry/131551.

Pennycook G and Rand DG (2021) The psychology of fake news. *Trends in cognitive sciences*, *25*(5): 388-402.

Pickles J (1995) *Ground Truth: The Social Implications of Geographical Information Systems*. Routledge.

Popper K (2002) *The Logic of Scientific Discovery*. Routledge.

Reuters Institute. *About the Reuters Institute*. Available at: https://reutersinstitute.politics.ox.ac.uk/about-reuters-institute.

Samek W and Müller KR (2019) Towards explainable artificial intelligence. 5-22. In Samek W, Montavon G, Vedaldi A, Hansen L K and Müller KR (eds) Explainable AI: interpreting, explaining and visualizing deep learning. *Lecture Notes in Computer Science*, vol 11700.

Scriven M (1953) The Mechanical Concept of Mind. *Mind*, 62(246), 230–240.

Searle JR (1980) Minds, brains, and programs. *Behavioral and brain sciences*, *3*(3): 417-424.

Stark PB (2018) Before reproducibility must come preproducibility. *Nature*, 557(7706): 613-614.

State L (2012) If It's Neutral, It's Not Technology *Educational Technology*, 52(1): 6-9.

Time (2018) Facebook Suspends Trump Election Data Firm for Policy Breach. Available at: https://time.com/5204387/facebook-suspends-cambridge-analytica.

Turing AM (1950) Computing Machinery and Intelligence. *Mind* 49: 433-460.

UN (2021) United Nations, Global Media and Information Literacy Week A/RES/75/267. Available at: https://documents-dds-ny.un.org/doc/UNDOC/GEN/N21/076/41/PDF/N2107641.pdf?OpenElement.

UK Government (2019) Digital, Culture, Media and Sports Committee. 'Disinformation and 'fake news': Final Report'. Available at: https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/1791.pdf.

Waisbord S (2018) Truth is What Happens to News: On journalism, fake news, and post-truth. *Journalism Studies*, 19(13): 1866–1878.

Wirtz BW, Weyerer JC and Sturm BJ (2020) The dark sides of artificial intelligence: An integrated AI governance framework for public administration. *International Journal of Public Administration*, 43(9): 818-829.

Wölker A and Powell TE (2021) Algorithms in the newsroom? News readers' perceived credibility and selection of automated journalism. *Journalism*, 22(1): 86–103.

[1] For example, the UCL Centre for Artificial Intelligence as a research centre with related programmes. https://www.ucl.ac.uk/ai-centre

[2] https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en

[3] https://digital-markets-act.ec.europa.eu/index_en

[4] Wikipedia s.v. Cambridge Analytica https://en.wikipedia.org/wiki/Cambridge_Analytica