Future Technology and Media Literacy:

Understanding Generative AI

Published 22 February 2024

Welsh overview available

Making Sense of Media

Contents

Section

Overview

Generative artificial intelligence ("generative AI") has rapidly gone from being a relatively unknown technology to a topic that has attracted significant mainstream attention from users, media and investors as the technology has become available for public use.

In the first two months of its launch in November 2022, the well-known generative AI model, ChatGPT, amassed over 100 million users. Analysts reported that, at its launch, it was the fastest-growing consumer internet app, comparing it with TikTok, which took nine months to reach 100 million users, and Instagram, which took over two years.[1] More broadly, USD 14.1 billion was invested in generative AI start-ups in the first six months of 2023 alone.[2]

Generative AI is being integrated into services such as social media, gaming, dating, applications (word-processing and spreadsheets) and search, among others: social media platform Snap has integrated 'MyAI', a conversational chatbot, into every users' app, which can engage with users to provide information, suggestions, and recommendations;[3] Microsoft has updated its search service Bing, so that generative AI now helps to provide a summary of live search results from across the web;[4] and the gaming platform Roblox has developed a tool that can generate new virtual worlds.[5]

> Ofcom's Online Nation report has found that 79% of 12-17 year olds are using generative AI tools and services.[6]

Many people have highlighted the potential of generative AI to transform our online experiences. Bill Gates has said that: 'The development of [generative] AI is as fundamental as the creation of the microprocessor, the personal computer, the Internet, and the mobile phone. It will change the way people work, learn, travel, get health care, and communicate with each other.'[7]

While generative AI could certainly provide new opportunities for people to learn, synthesise information at speed and generate content, it may also present new risks. Some of the opportunities and risks of generative AI will be familiar to those working to promote media literacy, as skills such as checking sources and critical engagement with information, and issues such as data protection and privacy, pre-date the internet and are relevant to our online lives today. However, generative AI is likely to cause some significant shifts in how these risks and opportunities arise and are experienced and will therefore require new applications of media literacy skills.

This document explores where these shifts may occur, what these opportunities and risks could be, and how platforms, the media literacy sector and users could respond.

As discussed at the UK government AI Safety Summit 2023, there are immediate threats that AI, including generative AI, could pose to international and national security as well as individual rights

[1] https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app

[2] https://www.cbinsights.com/research/generative-ai-funding-top-startups-investors/

[3] https://techcrunch.com/2023/04/19/snapchat-opens-its-ai-chatbot-to-global-users-says-the-ai-will-later-snap-you-back/

[4] https://www.reuters.com/technology/microsoft-infuse-software-with-more-ai-google-rivalry-heats-up-2023-02-07/

[5] https://aibusiness.com/ml/roblox-gets-generative-ai-users-can-build-virtual-worlds-from-text

[6] https://www.ofcom.org.uk/news-centre/2023/gen-z-driving-early-adoption-of-gen-ai

[7] https://www.gatesnotes.com/The-Age-of-AI-Has-Begun

and safety.[8] This paper does not address those issues but instead focuses on the media literacy implications of generative AI, particularly in relation to understanding and trust in online content.

Key areas this paper explores in relation to media literacy:
Accuracy of generative AI
Mis/disinformation
Amplification of bias

A future paper will look at the applications of generative AI including news, personalisation, creation, education, and data privacy and the implications of media literacy.

[8] https://www.aisafetysummit.gov.uk/

Introduction to the future technology trends project

This discussion paper forms part of a series Ofcom is producing on future technology trends and their related potential media literacy implications, in order to support those working on media literacy to better understand what opportunities and challenges could arise.

To outline Ofcom's role in this area, where this remit originates and how Ofcom defines media literacy, we have created an anchor document.[9] It describes how we will select the future technology trends that will be looked at in this discussion series and the lenses through which we will assess the media literacy implications of those trends.

[9] https://www.ofcom.org.uk/__data/assets/pdf_file/0025/263374/Anchor-Document.pdf

A definition of generative AI

Generative AI refers broadly to machine learning models that can create new content. Models create a wide variety of outputs including text, video, and audio.

Generative AI is usually based on 'foundation models' that are pre-trained through significant amounts of data to predict the missing parts of masked text or images from huge data sets. For example, the last time OpenAI shared ChatGPT's training size was with GPT-3[10], when the company said it was trained on 300 billion tokens (i.e. strings of words) at the time.[11] These models are then fine-tuned for purposes such as answering questions with conversational data and human feedback. This results in a model that can produce a plausible and understandable response, although there is no guarantee that the response is factually correct.

Foundation models can be used to power image generators, audio generators, code generators and large language models, which are models that can generate text.

Generative AI is already being used in a wide range of applications, from creative tasks like art and content generation to practical applications such as data augmentation,[12] summarising text, and language translation. It's an area of AI that continues to evolve and has the potential to support human creation and learning, as well as supporting technologies to improve inclusivity through its adaptive and language capabilities.

[10] https://www.techtarget.com/searchenterpriseai/definition/GPT-3
[11] https://www.cnbc.com/2023/05/16/googles-palm-2-uses-nearly-five-times-more-text-data-than-predecessor.html
[12] https://research.aimultiple.com/data-augmentation/

Opportunities and Risks

Like all new technologies, generative AI also comes with challenges including concerns about the accuracy of the generated content, as well as issues related to the spread of mis/disinformation, bias, privacy, and security. Although many of the risks associated with generative AI are similar to those posed by longstanding and less sophisticated forms of AI, they may be amplified given the volume and sophistication of the content generative AI models can create and the increase in their use by both companies and individuals.

As generative AI technologies advance and become more widely used, addressing these challenges will become increasingly important. Media literacy can help to address some of these risks, but it will likely not be a silver-bullet, particularly given the potential for malicious actors to exploit the features of generative AI to perpetrate harm. It may be that the level of media literacy skill and consistency of application needed to mitigate the risks of generative AI are not realistic for the majority of users.

As outlined in the anchor document for this series, we have used Ofcom's media literacy definition ('*the ability to use, understand and create media and communications in a variety of contexts*')[13] and the European Commission's Digital Competencies framework[14] (the "Framework") as a structure for examining the media literacy opportunities and risks that may arise in relation to generative AI. In this paper we have explored some of the areas where there is the potential for a significant shift in either the opportunity or risk (and sometimes both) that could arise from increased use and prevalence of generative AI.

As outlined in the Making Sense of Media Annual Plan 2023-24,[15] media literacy is about both people and platforms. The considerations below are designed to form part of a sector-wide discussion on how media literacy could be supported through improving people's skills and the actions of platforms.

[13] https://www.ofcom.org.uk/research-and-data/media-literacy-research

[14] Vuorikari, R., Kluzer, S. and Punie, Y., DigComp 2.2: The Digital Competence Framework for Citizens - With new examples of knowledge, skills and attitudes, Publications Office of the European Union, Luxembourg, 2022

[15] Ofcom, Making Sense of Media annual plan 2023-24, https://www.ofcom.org.uk/research-and-data/media-literacy-research/approach

Accuracy of generative AI content

The increasing prevalence of generative AI across online media means users are becoming more likely to encounter the technology (including content generated by it) in all aspects of their online lives. It can support content creation across various media, generating news articles, essays, web pages, marketing copy, social media posts, pictures, audio, and video (among others). It can also analyse articles, research papers and books and can be used to answer user questions as a chatbot. However, there are a number of challenges related to the accuracy of the content produced by generative AI tools, which present opportunities and risks for users' media literacy.

> Generative AI will often give responses to questions or prompts with definitive statements, but this is not a reflection of how the technology itself works. Generative AI is, at its base, a model trained on large sets of data that then predicts the most likely response to prompts or questions based on the patterns found in that data.

While this often produces results that are factually accurate, generative AI models have been known to produce 'hallucinations,' which are answers that have been generated as the probable correct answer based on the patterns in the training data but are in fact incorrect and sometimes completely nonsensical.[16] Recent research indicates that hallucinations in prominent generative AI models currently vary between 3 and 30%.[17]

These incorrect responses can be 'open domain,' when the generative AI tool is asked a question and gives a factually inaccurate response, or 'closed domain,' when a system is asked to summarise a document and adds information that is not present in the document.

Generative AI algorithms work in ways that are usually not visible or easily understood by users, making it difficult to understand why specific answers or responses have been given. This is often referred to as "black box" decision-making as it may be impossible to trace back how and why an algorithm makes specific suggestions or predictions and sometimes generates inaccurate or incorrect information. Some companies, such as Microsoft's 'Co-pilot' (previously Bing), have chosen to cite their sources of information, which will allow users to see the content in references.[18]

[16] https://www.techtarget.com/searchenterpriseai/definition/generative-AI
[17] https://www.nytimes.com/2023/11/06/technology/chatbots-hallucination-rates.html
[18] https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/

Accuracy of information: media literacy considerations

Do users understand that they are interacting with generative AI?

As the use of generative AI increases online it will likely become increasingly challenging for users to identify that they are engaging with generative AI tools or content. In a study conducted in 2023, linguistics experts were only able to identify research abstracts generated by AI 28.9% of the time.[19] When users are interacting with a chatbot, they may not be able to distinguish between a generative AI chatbot and a real human being, particularly as the language generated by these models can perceive and replicate tone to make them appear more 'human.'

Users would be better placed to understand they are interacting with generative AI if they were made aware that it is being utilised in the online space they are in, for example by clear signposting of the use of generative AI.

Are users aware that generative AI only produces probable answers?

In part because of the relatively recent use of generative AI among the public, and in part because many of the most popular generative AI models do not provide prominent explanations of how the technology works, it is not clear that users have the opportunity to fully understand how generative AI works, or that the answers it produces are not necessarily correct.[20] This creates a risk that users adopt factually incorrect information across their online and offline lives, a risk that is more significant as the numbers of generative AI users increase.[21] If these matters were flagged as users entered their prompts or questions, it may support people's understanding of the risk of incorrect answers.

How do users identify incorrect answers?

Generative AI can produce incorrect answers or hallucinations. However, users may rely on an answer produced by a generative AI tool as fact. In addition to understanding the technology, in order to assess whether an answer can be trusted, users will need to have high levels of media literacy skills to critically evaluate and assess the accuracy of information produced by generative AI. This is an existing media literacy skill that can be applied to users' interactions with generative AI.

[19] https://neurosciencenews.com/ai-human-writing-chatgpt-23892/

[20] https://www2.deloitte.com/uk/en/pages/press-releases/articles/more-than-four-million-people-in-the-uk-have-used-generative-ai-for-work-deloitte.html

[21] https://www2.deloitte.com/uk/en/pages/press-releases/articles/more-than-four-million-people-in-the-uk-have-used-generative-ai-for-work-deloitte.html

Users may not understand how content is generated

There are a range of possible scenarios where it may become more difficult for users to understand how the content they are consuming has been created. News literacy puts the emphasis on audiences understanding the work that journalists put into creating content, from information gathering to fact checking to editing for balance. It may become harder for users to trust the content they see as generative AI is introduced into the news writing process, meaning users do not have the name of a reporter or an opportunity to understand exactly how a story has been produced.

Mis/disinformation and synthetic media

One of the primary concerns arising from the advent of generative AI is the impact it may have on the spread of mis/disinformation.

The UK government defines disinformation as the deliberate creation and spreading of false and/or manipulated information that is intended to deceive and mislead people, either for the purposes of causing harm, or for political, personal or financial gain. Misinformation is the inadvertent spread of false information.[22]

Generative AI can be used to create mis/disinformation content that mimics authentic information sources and is therefore more believable.

> As generative AI becomes increasingly convincing at mimicking human-generated content, it is likely that the lines between real and fake content will become increasingly blurred, making it more challenging for users to discern what information is authentic and what is AI generated mis/disinformation.

A focal point when considering the risks posed by generative AI has been the development of synthetic media (i.e., media produced by generative AI), particularly 'deepfakes.' Deepfakes are a type of synthetic media. Deepfakes can be defined as content that has been manipulated or created outright using AI or related digital techniques, with the explicit purpose of deceiving audiences.

The ability to produce audio, video, and image mis/disinformation has in the past been more limited, for example by digitally altering content (e.g., slowing down or cutting). These methods are far less sophisticated, and it is often possible to identify this content as having been edited.[23] However, synthetic media and deepfakes are becoming more sophisticated as generative AI tools become cheaper, more accessible, and easier to command and use. This AI generated synthetic media can include images, audio, code, and video, among others. While these tools can be used for user entertainment,[24] they also introduce a tool for those seeking to alter public opinion and undermine the credibility of authentic sources of information, including in important public and political conversations and debates. Synthetic media and deepfakes can replicate the likeness of trusted individuals, making inflammatory statements, propagating mis/disinformation, or generally sowing uncertainty and mistrust. The Financial Times reported that deepfake 'news' videos were used to spread misinformation in Venezuela[25]. A UK tech company, Synthesia, was used to produce fake news widely circulated in media supportive of the government.[26] It is key that users understand what media is authentic and what is AI generated. If they do not, there is a significant risk of

[22] https://www.gov.uk/government/news/fact-sheet-on-the-cdu-and-rru
[23] https://www.theguardian.com/us-news/video/2019/may/24/real-v-fake-debunking-the-drunk-nancy-pelosi-footage-video
[24] https://www.forbes.com/sites/danidiplacido/2023/03/27/why-did-balenciaga-pope-go-viral/?sh=38ec4f5e4972
[25] https://www.ft.com/content/3a2b3d54-0954-443e-adef-073a4831cdbd
[26] https://www.ft.com/content/3a2b3d54-0954-443e-adef-073a4831cdbd

undermining trust in all content, as users may conclude they should distrust everything they see online rather than risk being taken in by mis/disinformation.

Generative AI's capacity to generate content tailored to specific demographics and interests provides the means for malicious actors to craft highly effective mis/disinformation campaigns. By leveraging this technology, mis/disinformation can be personalised in an attempt to resonate with particular groups of people or subsets of more vulnerable users, increasing the likelihood of these users engaging with that content and therefore deepening the challenge of countering mis/disinformation. This is compounded by the ability of generative AI to create content in multiple languages, enabling malicious actors to reach a wider audience more quickly and easily.

AI generated content can be difficult for individuals to verify through conventional means. To assist users, larger online platforms and news sources have taken steps to verify the factual accuracy of content, taking down inaccurate information, labelling misleading content and offering fact checking. However, generative AI can make it simpler to create a paper trail of supporting 'evidence' to underpin mis/disinformation content, making it harder for users to verify the accuracy of the information in question. For example, malicious actors could develop deepfakes of trusted individuals and plant a trail of AI generated content from what appear to be legitimate sources to provide 'evidence' to reinforce the mis/disinformation.[27]

It may therefore become increasingly difficult for users and online platforms to adequately identify and address AI generated mis/disinformation, despite the efforts of large platforms to address this problem on their services. This is also a challenge for trusted media sources, as the speed and quantity at which malicious actors can produce mis/disinformation can make it harder for them to verify and challenge information before it is spread.

Mis/disinformation: media literacy considerations

Mis/disinformation at scale

As discussed above, the development of more sophisticated ways of presenting mis/disinformation to users, such as deepfake videos of trusted figures, voice replication etc. may make it harder for users to identify mis/disinformation. It will be important that users are aware of the risks of synthetic media when verifying this information.

While news media organisations are looking to fund ways to verify content,[28] users will need to be aware of the potential of deepfakes and synthetic media. They will also need to have strong media literacy skills to support them to discern legitimate sources and have a range of trusted sources to enable them to verify information they see online. Users could be assisted in this through the development of verified lines of communication from high profile individuals, news, and information sources to ensure they have access to high quality information. These media literacy skills and assistance from high quality and trusted news sources will be key in ensuring that users continue to trust high

[27] https://time.com/6216722/how-ai-tech-harms-children/

[28] https://c2pa.org/

quality information and news, rather than instead mistrusting all information they see online.

Generative AI amplification of mis/disinformation

Generative AI has the potential to learn and reproduce mis/disinformation as if it were true. These models are trained on vast datasets from the internet, and the quality and accuracy of the information in those datasets can vary. If a model is exposed to a significant amount of false or misleading information during its training, it may generate content that replicates this mis/disinformation. It's important to note that generative AI models are not capable of determining the truth or accuracy of information on their own. They do not possess an inherent understanding of facts or the ability to verify the information they generate. They simply mimic patterns in the training data. For instance, in 2020, an investigation by the Center on Terrorism, Extremism, and Counterterrorism found that GPT-3 could be prompted to perform tasks such as producing discourse reminiscent of the New Zealand Christchurch shooter, reproducing fake forum threads casually discussing genocide and promoting Nazism in the style of the defunct Iron March community.[29]

It will be important for training data for generative-AI models to be closely monitored, to minimise the amount of mis/disinformation these models are trained on, although high levels of oversight becomes more difficult as the size of the datasets that models are trained on get bigger and bigger. Users will need to critically evaluate the credibility of data, information and digital content produced by generative AI and remain vigilant to the possibility of mis/disinformation being reproduced through generative AI.

Sources of content

The inability to understand how content is generated also creates a challenge when determining the source of content. Historically, knowing where content comes from enables audiences to decide whether it has come from a trusted source. Being able to evaluate and fact check online information is a key news literacy skill that helps audiences identify mis/disinformation. The black box nature of generative AI poses problems for tried and tested fact-checking approaches, although there are also future possibilities that generative AI could be used to aid fact checking with organisations such as Full Fact are working to develop AI systems that are able to auto fact check content.[30] These tools and techniques to identify AI generated content are currently

[29] https://www.middlebury.edu/institute/academics/centers-initiatives/ctec/ctec-publications/radicalization-risks-gpt-3-and-neural-language

[30] https://fullfact.org/about/ai/

developing, are relatively small, and are not completely effective,[31] so even platforms and news organisations themselves may struggle to identify AI generated content.

[31] https://www.gmfus.org/news/ai-startups-and-fight-against-misdisinformation-online-update

Amplification of bias

The vast datasets on which generative AI models are trained may contain biased or prejudiced information which generative AI could amplify. Whilst this is reflective of human bias already present in society, the combination of training data with bias and the design of algorithms means that bias will be reproduced across the outputs of generative AI models. For example, if the training data contains biased language or narratives, such as harmful stereotypes about certain groups, the model may learn and replicate these biases in the content it generates, reinforcing and perpetuating pre-existing societal biases and prejudices. The risks of training data containing and reproducing undesirable or even illegal content has been highlighted by the recent finding that a prominent image training dataset called LAION-5B contained child sexual abuse material[32].

> Generative AI models can underrepresent or exclude the voices and perspectives of marginalised or minority groups where they are underrepresented in the training data. This can lead to the production of inaccurate information, underrepresentation, or exclusion of information about those groups, perpetuating the existing biases and lack of diversity found in the training data.

Equally, it is possible for generative AI models to overrepresent certain groups or topics if they are overrepresented in the training data, which can lead to content that is disproportionately focused on those groups or topics.

It is also possible for generative AI to produce content with confirmation bias. If a user provides a biased prompt, for example a prompt that includes a clear political or moral bias, the model may produce content that confirms or amplifies those biases. This can occur due to the interplay between the user's input, the model's training data, and the AI's propensity to generate content that is contextually relevant. This can mean that users who exhibit bias in their prompts may only receive information or content that aligns with or reinforces that bias, decreasing the diversity and plurality of the information consumed by that user and limiting their exposure to alternative viewpoints.

[32] https://www.telegraph.co.uk/business/2023/12/20/fears-ai-trained-child-abuse-images-thousands-discovered/

Amplification of bias: media literacy considerations

Over- or underrepresentation of marginalised groups

Overrepresentation or underrepresentation of certain groups in training data can lead to AI generated content that is biased, unfair, and unrepresentative of diverse perspectives, potentially perpetuating existing inequalities, prejudices, and discrimination. Underrepresentation of certain groups in training data can lead to these groups being marginalised and ignored in AI generated content, meaning they may not find content that speaks to their experiences or needs.

For example, when prompted to develop images of black, or African American women, some generative AI models have failed to produce realistic images, with the AI company stating that this is caused by "overrepresentation" of white people in its general data sets.[33]

This exclusion can further marginalise underrepresented communities, making them feel disregarded and underrepresented in the digital world. It can hinder inclusivity and prevent diverse voices from being heard.

To mitigate this risk the data sets generative AI models are trained on should be diverse and representative. This may require greater human oversight as well as bias testing once the models are trained to reduce the generation of biased content.

Reinforcement of stereotypes

Generative AI models learn from the data they are trained on, which may contain stereotypes or misconceptions about certain groups. These stereotypes could be explicit or subtle biases present in the training data, including text, images, or other forms of media. As AI models learn to generate content, they may inadvertently reproduce and amplify these stereotypes. For example, if a generative AI model is trained on a dataset that contains a bias against a particular ethnic group, it may produce content that reflects these biases in the form of biased language, discriminatory assumptions, or negative portrayals. AI generated content that reinforces stereotypes perpetuate harmful narratives about certain groups, often marginalising or devaluing them.[34] This can lead to social injustice and contribute to systemic discrimination. For example, Bloomberg research found that image sets generated for every high-paying job were dominated by subjects with lighter skin tones, while subjects with darker skin tones were more commonly generated by prompts like "fast-food worker".[35] As stated by Heather Hiles, chair of Black Girls Code, "People learn

[33] https://www.nytimes.com/2023/07/04/arts/design/black-artists-bias-ai.html
[34] https://counterhate.com/research/misinformation-on-bard-google-ai-chat/
[35] https://www.bloomberg.com/graphics/2023-generative-ai-bias/

from seeing or not seeing themselves that maybe they don't belong. These things are reinforced through images."

If AI generated content which contains stereotypes is disseminated widely online, it can lead to the spread of misinformation and biased perspectives about certain groups, negatively affecting public discourse. Additionally, the continuous exposure to biased content generated by AI can influence individuals' perceptions, affecting their understanding of different groups. This can lead to discrimination, misunderstanding, and even hostility in both online and offline interactions.

To minimise the reproduction of stereotypes it is important the generative AI models are trained on diverse and unbiased training data: ensuring that training data is diverse, representative, and free from explicit biases can reduce the risk of the model learning and reproducing stereotypes. While there are significant questions as to whether any bias-free data sets exist, companies could check for and attempt to cleanse training data of bias. There are also bias mitigation techniques that could be implemented during model training or post-training to reduce the generation of content that reflects stereotypes.

It is also important that users are aware of the possibility of the reinforcement of stereotypes. This requires users to question the content produced by generative AI, including the language used and the portrayals of individuals or groups of people. It also involves users questioning the stereotypes they hold and their own biased opinions about other people or groups of people (whether consciously or unconsciously held), removing the biases they hold offline to enable them to question the content generated online.

Ultimately, many of the issues of amplification and production of bias in generative AI is often at its root a human problem. It is human data, systems and stereotypes that are often biased, with generative AI amplifying these. While mitigations can be put in place, the ultimate mitigation would be tackling human biases.

Conclusion

Generative AI is changing the way we engage with information online While generative AI is already in use, it is not yet clear what the lasting impact will be on users and society, and how it will shape our behaviours in the future.

A new way of engaging with online content does not necessarily mean completely new media literacy skills are needed. Many of the skills referred to throughout this discussion paper are skills that are already needed to navigate the internet today, but with different applications and skill levels needed in increasingly complex environments.

Understanding and shaping how generative AI will change our world, and what it means for media literacy, will be an important task for the years ahead and so these questions may provoke fruitful discussion:

Questions for discussion

1. Does the responsibility for media literacy remain with platforms, professionals, and parents as well as users themselves or are there other actors with responsibility too?
2. How are users supported to recognise when generative AI has been used and the limits and quality of the information it produces?
3. What are users' attitudes towards generative AI? To what extent are users critically engaging with generative AI, and does that differ across age, gender, and socio-economic backgrounds?
4. What is it reasonable to expect of users and where are the boundaries of media literacy compared to technological innovation and regulation?