Cambridge Consultants

USE OF AI IN ONLINE CONTENT MODERATION

2019 REPORT PRODUCED ON BEHALF OF Ofcom

Foreword from Ofcom

As the UK's converged communications regulator, Ofcom oversees broadband and mobile telecoms, TV, radio, video-on-demand services, post and the airwaves used by wireless devices. We have a statutory duty to promote media literacy, under which we carry out research into people's use of – and attitudes towards – various communications services, including online services such as social media and video sharing platforms. This remit reflects the increasing convergence of the companies and markets that we regulate.

In recent years a wide-ranging, global debate has emerged around the risks faced by internet users, with a specific focus on protecting users from harmful content. A key element of this debate has centred on the role and capabilities of automated approaches (driven by Artificial Intelligence and Machine Learning techniques) to enhance the effectiveness of online content moderation and offer users greater protection from potentially harmful material. These approaches may have implications for people's future use of – and attitudes towards – online communications services, and may be applicable more broadly to developing new techniques for moderating and cataloguing content in the broadcast and audiovisual media industries, as well as back-office support functions in the telecoms, media and postal sectors.

Ofcom has commissioned Cambridge Consultants to produce this report as a contribution to the evidence base on people's use of and attitudes towards online services, which helps enable wider debate on the risks faced by internet users.

EXECUTIVE SUMMARY

In the last two decades, online platforms that permit users to interact and upload content for others to view have become integral to the lives of many people and have provided a benefit to society. However, there is growing awareness amongst the public, businesses and policy makers of the potential damage caused by harmful online material. The user generated content (UGC) posted by users contributes to the richness and variety of content on the internet but is not subject to the editorial controls associated with traditional media. This enables some users to post content which could harm others, particularly children or vulnerable people. Examples of this include content which is cruel and insensitive to others, which promotes terrorism or depicts child abuse.

As the amount of UGC that platform users upload continues to accelerate,[1] it has become impossible to identify and remove harmful content using traditional human-led moderation approaches at the speed and scale necessary.

This paper examines the capabilities of artificial intelligence (AI) technologies in meeting the challenges of moderating online content and how improvements are likely to enhance those capabilities over approximately the next five years.

Recent advances in AI and its potential future impact

The term 'AI' is used in this report to refer to the capability of a machine to exhibit human-like performance at a defined task, rather than refer to specific technical approaches, such as 'machine learning'. Although AI has been through several cycles of hype followed by disillusionment since its inception in the 1950s, the current surge in investment and technological progress is likely to be sustained. The recent advances in AI have been enabled by progress in new algorithms and the availability of computational power and data, and AI capabilities are now delivering real commercial value across a range of sectors.

The latest advances in AI have mainly been driven by machine learning, which enables a computer system to make decisions and predict outcomes without being explicitly programmed to perform these tasks. This approach requires a set of data to train the system or a training environment in which the system can experiment. The most significant breakthrough of machine learning in recent times is the development of 'deep neural networks' which enable 'deep learning'. These neural networks enable systems to recognise features in complex data inputs such as human speech, images and text. For many applications, the performance of these systems in delivering the specific task for which they have been trained now compares favourably with humans, but AI still does make errors.

The advances of AI in recent years will continue, driven by commercial applications and enabled by continued progress in algorithm development, the increasing availability of low-cost computational power and the widespread collection of data. There are, however, some inhibitors to making the most of the potential of AI, such as the lack of transparency of some AI algorithms which are unable to fully explain the reasoning for their decisions. Society has not yet built up the same level of trust in AI systems as in humans when making complex decisions. There are risks that bias is introduced into AI systems by incorporating data which is unrepresentative or by programming in the unconscious bias of human developers. In addition, there is a shortage of staff suitably qualified in developing and implementing AI systems. However, many of these inhibitors are being addressed: the supply of AI-skilled engineers will increase and it is likely that society will gradually develop greater confidence in AI as it is increasingly seen to perform complex tasks well and it is adopted in more aspects of our lives. Other issues will, however, continue to be a problem for at least the short term and these are discussed in more detail in this report.

Current approaches and challenges to online content moderation

Effective moderation of harmful online content is a challenging problem for many reasons. While many of these challenges affect both human and automated moderation systems, some are especially challenging for AI-based automation systems to overcome.

There is a broad range of content which is potentially harmful, including but not limited to: child abuse material, violent and extreme content, hate speech, graphic content, sexual content, cruel and insensitive material and spam content. Some harmful content can be identified by analysing the content alone, but other content requires an understanding of the context around it to determine whether or not it is harmful. Interpreting this context consistently is challenging for both human and automated systems because it requires a broader understanding of societal, cultural, historical and political factors. Some of these contextual considerations vary around the world due to differences in national laws and what societies deem acceptable. Content moderation processes must therefore be contextually aware and culturally-specific to be effective.

Online content may appear in numerous different formats which are more difficult to analyse and moderate, such as video content (which requires image analysis over multiple frames to be combined with audio analysis) and memes (which require a combination of text and image analysis with contextual and cultural understanding). Deepfakes, which are created using machine learning to generate fake but convincing images, video, audio or text, have the potential to be extremely harmful and are difficult to detect by human or AI methods.

In addition, content may be posted as a live video stream or live text chat which must be analysed and moderated in real time. This is more challenging because the level of harmfulness can escalate quickly and only the previous and current elements of the content are available for consideration.

Over time, the language and format of online content evolve rapidly and some users will attempt to subvert moderation systems such as by adjusting the words and phrases they use. Moderation systems must therefore adapt to keep pace with these changes.

Online platforms may moderate the third-party content they host to reduce the risk of exposing their users to harmful material and to mitigate the reputational risk to the organisation. However, removing content which is not universally agreed to be harmful can also result in reputational damage and undermine users' freedom of expression. Facebook, for example, has been criticised for removing an image of a statue of Neptune in Bologna, Italy for being sexually explicit and the iconic photograph of a young girl fleeing a napalm bombing during the Vietnam War for showing child nudity.

The variety of types of harmful content makes it difficult for online platforms to define in their community standards the content and behaviours which are not permitted on their platforms. AI-enabled content moderation systems are developed to identify harmful content by following rules and interpreting many different examples of content which is and is not harmful. It can be challenging for automated systems to interpret the community standards of a platform to determine whether content is harmful or not. Clarity in platforms' community standards is therefore essential to enabling the development and refinement of AI systems to enforce the standards consistently.

Overall, it is not possible to fully automate effective content moderation. For the foreseeable future the input of human moderators will continue to be required to review highly contextual, nuanced content. Human input is expensive and difficult to scale as the volume of content uploaded increases. It also requires individuals to view harmful content in order to moderate it, which can cause them significant psychological distress. However, AI-based content moderation systems can reduce the need for human moderation and reduce the impact on them of viewing harmful content.

Many organisations follow an online content moderation workflow which uses moderation at one or both of two points in the workflow:

1. **Pre-moderation** when the uploaded content is moderated prior to publication, typically using automated systems

2. **Post-** or **reactive-moderation** when content is moderated after it has been published and it has been flagged by other users or automated processes as potentially harmful, or which was removed previously but requires a second review upon appeal

This example workflow is shown in Figure 1 below, which also indicates where AI can be used beneficially.

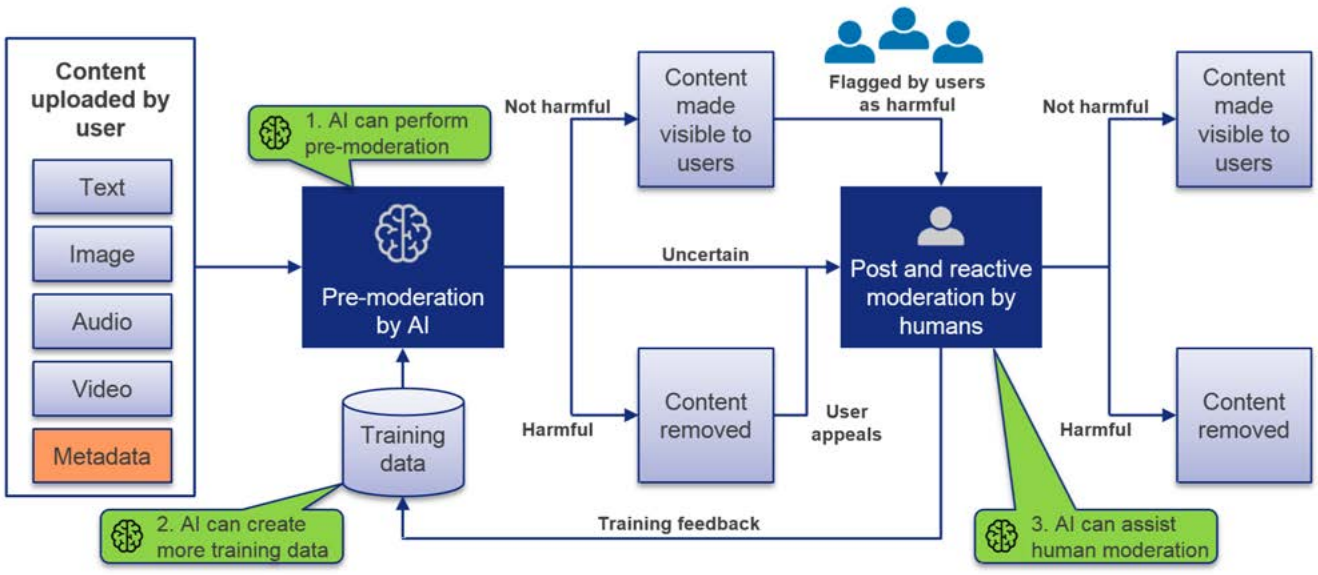


Figure 1 – There are three key ways in which AI can improve the effectiveness of the typical online content moderation workflow (**SOURCE:** Cambridge Consultants)

The potential impact of AI in online content moderation

AI technologies have the potential to have a significant impact on the content moderation workflow in three ways:

1. **AI can be used to improve the pre-moderation stage and flag content for review by humans, increasing moderation accuracy**
   Relatively simple approaches such as 'hash matching', in which a fingerprint of an image is compared with a database of known harmful images, and 'keyword filtering', in which words that indicate potentially harmful content are used to flag content, are useful tools but have limitations. Detecting the nuances of language, such as sarcasm and the use of emojis, is difficult, and languages - and in particular slang terms - evolve over time. 'Natural language understanding'[2] and 'sentiment analysis'[3] techniques are a focus of research and are becoming increasingly effective. Similarly, 'object detection'[4] and 'scene understanding'[5] are essential capabilities for moderating complex image and video content and these capabilities have advanced in recent years. An AI approach known as 'recurrent neural networks'[6] can enable more sophisticated analysis of video content, which is particularly challenging to moderate as frames must be considered relative to other frames in the video.

   AI moderation techniques can also increasingly consider the context in which content appears, although in general this remains complex and challenging. In practice, most harmful content is generated by a minority of users and so AI techniques can be used to identify malicious users and prioritise their content for review. 'Metadata' encodes some context relevant to moderation decisions about content, such as a user's history on the site, the number of friends or followers they have and information about the user's real identity, such as age or location. There is also a cultural and historical context in many online interactions. Any preceding content, such as previous interactions between individual users or the flow of the discussion, can provide valuable context which can be analysed alongside the content itself. The metadata available varies between platforms and for the type of content posted and so it is difficult for platform-agnostic moderation tools to take full advantage of metadata when making moderation decisions.

   Different AI architectures are required for identifying different categories of potentially harmful content. For example, identifying child abuse material requires consideration of the content (an image or video) but in general the context (such as the age or location of the user posting it or the number of followers they have) is not an important factor in detecting it automatically. AI techniques such as 'object detection' and 'scene understanding' are essential elements of automated systems to identify this type of material. On the other hand, identifying bullying content often requires full consideration of the context of the user interactions as well as the content itself as the characteristics of bullying content are less well defined. The complexity of designing AI architectures for moderating different categories of content therefore increases the costs and challenges for organisations to develop these.

2. **AI can be implemented to synthesise training data to improve pre-moderation performance**
   Generative AI techniques, such as 'generative adversarial networks' (GANs) can create new and original images, video, audio or text. These approaches can be used to create images of harmful content such as nudity or violence. These images can supplement existing examples of harmful content when training an AI-based moderation system.

   To illustrate the capabilities of this technology, Figure 2 shows some faces which have been created by a GAN to resemble celebrities but are not faces of real people. This approach could be used to create images of harmful content while preserving the anonymity of victims shown in real examples of harmful content. GANs can also apply 'style transfer', a technique to change the style of one image into the style of another. Examples of this are shown in Figure 3 opposite. Style transfer is particularly valuable in correcting for bias in datasets by generating content of an under-represented minority. Generating more data out of a limited dataset is a valuable approach for augmenting training datasets when developing AI-based moderation systems. This permits them to make more accurate decisions and reduces their dependence on the availability of large datasets containing suitably anonymised content.

Figure 2 – Celebrity faces which are not of real people, but appear to be credible, have been created using GANs (**SOURCE:** Nvidia,[7] used with permission)

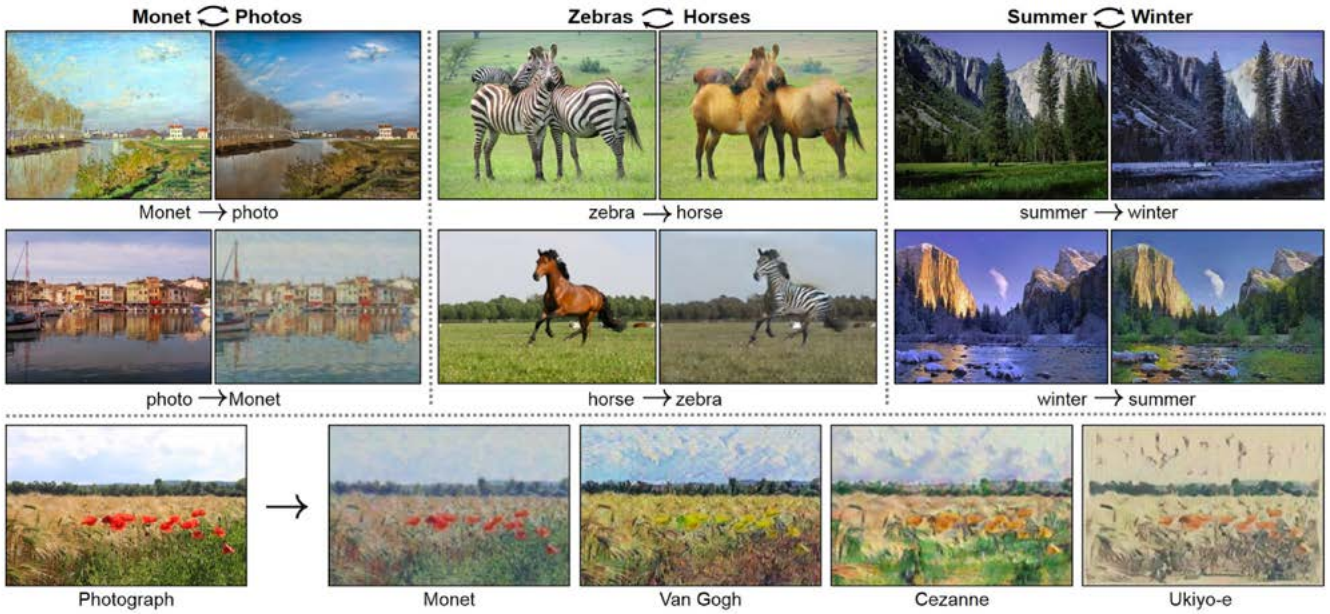


Figure 3 – GANs can apply style transfer to create new images in the style of others (**SOURCE:** Berkeley AI Research Laboratory, UC Berkeley,[8] used with permission)

3. **AI can assist human moderators by increasing their productivity and reducing the potentially harmful effects of content moderation on individual moderators**
   AI can improve the effectiveness of human moderators by prioritising content to be reviewed by them based on the level of harmfulness perceived in the content or the level of uncertainty from an automated moderation stage. It can reduce the impact on human moderators by varying the level and type of harmful content they are exposed to. It can limit exposure to the most harmful elements of the content, such as by identifying and blurring out areas of images which the moderator can optionally view only if needed to make a moderation decision. An AI technique known as 'visual question answering' allows humans to ask the system questions about the content to determine its degree of harmfulness without viewing it directly. This can reduce the harmful effects on human moderators but is less reliable than when human moderators do view the content directly. AI can also reduce the challenges of moderating content in different languages by providing high quality translations. Therefore, the productivity of human moderators can be increased and the harmful effects of viewing content can be reduced.

Commercial challenges for organisations using AI techniques

There are commercial challenges to organisations using AI techniques for online content moderation. The fast-paced, highly-competitive nature of online platforms can drive businesses to prioritise growing an active user base over the moderation of online content. Developing and implementing an effective content moderation system takes time, effort and finance, each of which may be a constraint on a rapidly growing platform in a competitive marketplace. Implementing AI technologies requires developers with the required skillset and who are much sought after within the technology industry and are therefore difficult and expensive to recruit.[9] AI-enabled moderation systems require access to suitable datasets for training, and organisations which have not already collected suitable data will need to acquire this either by gathering it themselves or by purchasing it from others.

Some barriers to developing AI-enabled content moderation tools may be greater for smaller organisations: such as access to skilled AI developers, datasets, financial resources and the greater impact of delayed platform development and growth. Therefore, smaller organisations may be less able to realise the benefits of advanced AI-enabled content moderation tools without intervention. There is a growing content moderation services sector[10] that appears to offer solutions to a range of sites and services, including small players. However, intervention may be needed if these services do not adequately address the barriers to smaller organisations accessing high-performance content moderation tools.

Many of the most successful internet platforms have evolved as content-sharing platforms, and profitability in many business models comes only after a critical mass of users and content

has been achieved.[11] This can incentivise the promotion of content which is 'clickable' and gets the attention of users by being sensational and in some cases may be harmful. There are therefore both incurred costs and opportunity costs for organisations investing in automated content moderation systems, but the reputation of organisations may be damaged if they do not moderate content appropriately.[12] Pressure to improve content moderation is growing due to a wider backlash against many of the larger internet companies and there are on-going public and policy debates about how this can be addressed.[13]

The potential impact of AI on promoting socially-positive online engagement

As well as actively moderating harmful content, AI technologies can be used to encourage positive engagement and discourage users from posting potentially harmful content in the first place.

The 'online disinhibition effect'[14] is a potential explanation for why some internet users act maliciously online. It suggests that multiple factors combine to drive malicious online behaviour. These factors include the anonymity of online interactions and the empathy deficit that this can foster. Most online communications are asynchronous, meaning the recipient does not necessarily receive the message and respond immediately, unlike face-to-face or telephone conversations. Hence users do not see any negative emotional reactions from others at that moment in time and may therefore communicate in a less inhibited manner.

Some non-AI approaches are used to help address the disinhibition effect: for example, multi-factor authentication of users to reduce their anonymity, demonetising potentially harmful content and penalising users by temporarily restricting privileges. AI techniques can similarly be used to encourage positive engagement online. For example, they can be used to inform users about content before they see it by using 'automatic keyword extraction' to detect meaning or sentiment. AI nudging techniques can discourage users from posting potentially harmful content by prompting them to think again or enforcing a short delay before content is posted if potentially harmful content is detected. Users can also be 'nudged' to post less negative content by suggesting alternatives that still represent the content of their original message. AI-powered chatbots which prompt users to report the harmful content of others have been shown to be effective and can encourage users to take more responsibility for their own posts. Similarly, chatbots posing as other users can be used to highlight negative content to those posting it and prompt them to improve their behaviour. An example of this approach is shown in Figure 4 below. These nudging techniques may be useful for reducing harmful interactions, but careful consideration is required to ensure they are implemented appropriately and ethically.

Figure 4 – Chatbot replies to racist messages can reduce the subsequent number of racist messages posted (**SOURCE:** Kevin Munger,[15] used with permission)

Policy implications for stakeholders

This report highlights four potential policy implications. These are to inform stakeholders from government, regulators, internet sites and services, and stakeholder groups of some of the possible areas of concern around AI-based content moderation.

These implications should be considered in the broader context of the evolving landscape of content moderation generally. Online sites and services offering transparency on their definition and treatment of harmful content will help to protect users and ensure that they, and other stakeholders, will gain confidence in the standards of the platform.

There is also promising research into the impact of techniques used to encourage socially positive online engagement. Should such techniques prove to be widely applicable, AI approaches can be used to discourage users from posting harmful content. This could reduce the reliance on online content moderation systems, given the current limitations of the technology, by reducing the amount of harmful content posted for some categories.

	CONTEXT	POLICY IMPLICATIONS
1	AI-enabled online content moderation tools are proving to be a necessary part of online platform providers' response to harmful online content and tools must be accessible to organisations of all sizes. Data is required for training AI systems, which could act as a barrier to some organisations.	The availability of online content moderation services from third-party providers should be encouraged. This will help to ensure that services are accessible to platforms of all sizes and will encourage the use of AI and automation techniques to increase the performance and effectiveness of content moderation.
2	The data used to train AI-enabled online content moderation services will evolve as different user behaviours come to the fore. Sharing data between organisations would assist the entire sector in adopting a collaborative and coordinated approach to managing harmful content.	The sharing of datasets applicable to identifying harmful content between platforms and moderation service providers will help to maintain standards and should be encouraged. Data Trusts (as identified in the UK Government's recent AI review[16]) may offer a suitable framework for this. Relevant data held by public bodies, such as the BBC, could be contributed for the benefit of society by enabling a comprehensive dataset to be made available and kept up to date with evolving categories of harmful content and formats.
3	Users and other stakeholders will need to feel confident in the performance of any AI-based content moderation approach, especially in the early days of technology adoption.	It is important to build public confidence that any potential sources of bias in AI-based content moderation are understood and appropriate steps taken to mitigate them. This may be through auditing and calibrating datasets to understand how representative they are of the diversity of individuals in society; or by establishing a testing regime for AI-based content moderation systems.
4	It is not always possible to fully understand how an AI system makes its decisions or how effective it is in meeting users' expectations of protection.	To ensure appropriate levels of internet user protection, it is important to be able to understand the performance of AI-based content moderation by individual platforms and moderation services across categories. This is to confirm that these deliver appropriate moderation and that they are evolving with the expectations of society and relevant national or cultural sensitivities.

TABLE OF CONTENTS

1 INTRODUCTION

The internet is highly beneficial to society but there is a growing awareness of the damage of harmful material online

Over recent decades the internet has become an integral part of almost everyone's life in the UK. Internet services now affect many different aspects of how we live our lives, from communicating with friends and family, accessing news and entertainment, to creating and publishing our own content and consuming content created by others. Overall, the internet is a major public benefit.[17] It has both a positive personal impact in terms of access to information and the ability to communicate with others, and a significant positive economic impact from the benefits it provides to organisations across all sectors of the economy.

However, the huge amounts of data being created on a daily basis, combined with growing concerns around the potential negative influences it has on individuals has triggered wider discussion amongst the public, policy-makers and sector stakeholders. In October 2017, the UK Government's Department for Digital, Culture, Media & Sport (DCMS) published[18] its "Internet Safety Strategy Green Paper" and launched a consultation[19] on the views of stakeholders. In April 2019, the UK Government published its "Online Harms White Paper" which sets out its plans for a package of measures to keep UK users safe online, including legislative measures to make companies more responsible for their users' safety online.

Harmful content includes content which is illegal, such as child abuse material or content promoting terrorist acts, through to material which is itself legal but still potentially harmful in the context of the viewers, such as images depicting violence or material designed to bully individuals. Determining the point at which to classify material as harmful can be challenging as it depends on the level of acceptability for the intended audience and historical or cultural contexts. Society is adjusting to the balance that is required between a right to freedom of speech and reasonable protection of the public to harmful content online.

There is a growing awareness within internet companies of their responsibilities in protecting the public and preventing their platforms from being used for undesirable, socially-negative purposes. The techniques used to identify and remove harmful material have generally relied on labour-intensive inspections by people, which is time-consuming, expensive, subject to variation and bias, not scalable and can impact the psychological wellbeing of those performing this task.

AI capabilities are advancing rapidly, enabled by new technologies such as deep learning

Rapid advances over the past decade in AI techniques have started to unlock the potential of the huge amounts of data which is now routinely collected and analysed. Improvements in computing power, in particular the use of graphical processing units (GPUs) which specialise in processing data in parallel, other chips designed specifically to execute AI algorithms and the availability of processing power in many end devices such as smartphones have enabled much of this progress.

Deep learning is a technique which employs multiple layers of neural networks which are trained on datasets to 'learn' specific characteristics of that data and then apply this learning to new data. This approach has been the subject of significant research and has enabled many of the high-profile breakthroughs in AI in recent years. One such breakthrough came in 2016 when Google DeepMind developed AlphaGo to play the game Go. It beat Lee Sedol, one of the world's leading players of the game. Just a few years before, it was expected that AI would not be sufficiently powerful to achieve such feats for some decades.

For the purposes of this report, we use the term artificial intelligence (AI) for technology which has the capability to exhibit human-like behaviour when faced with a specific task. Terms such as machine learning and deep learning refer to specific techniques in which the technology 'learns' how to provide AI capability and we therefore use these terms intentionally to refer to these techniques. As we use AI to describe the output or capability, we include within it some human-designed algorithms which may output a relatively simple function of the inputs, which some in the industry would exclude from the definition of AI.

Artificial intelligence shows promise in moderating online content but raises some issues

As AI techniques are very well suited to rapidly processing data and spotting patterns, they are ideal for being part of the solution to the problems of moderating online content. The technical scope of addressing content in text, images, video and audio is challenging, and human-like understanding of these media is required in many instances to identify harmful content. Recent advances in natural language understanding, sentiment analysis and image processing are key to enabling effective online content moderation at the scale which is required in the modern world.

However, there are also some issues in using AI for this purpose such as unintentional bias, a lack of transparency or 'explainability' in how decisions are made and how accuracy, speed and the availability of training data can be optimised. AI is not a silver bullet, and even where it can be successfully applied, there will be weaknesses which will be exploited by others to subvert the moderation system.

This study examines how AI technologies can have an impact on the moderation of online content

Ofcom has asked Cambridge Consultants to research and evaluate the potential impact of AI on the moderation of online content. We have used our deep expertise in the development of AI systems and our broad sector knowledge, coupled with structured interviews with key stakeholders within the industry, to produce this report.

The purpose of this report is to examine how the current state-of-the-art AI technologies can be used to improve the moderation of online content and how its capabilities are expected to evolve in the near future. This is to inform the wider debate being held by stakeholders on the impact of harmful content online and the steps which can be taken to reduce the potential for harm, while safeguarding considerations such as the freedom of speech.

Cambridge Consultants has retained editorial responsibility for this report throughout the process and the opinions stated are not necessarily those held by Ofcom. The highlighted policy implications made by Cambridge Consultants in this report are intended to inform stakeholders from government, regulators, internet sites and services, and stakeholder groups of some of the potential emerging challenges should AI-based content moderation become a mainstay of online content moderation.

This report takes a structured approach in considering the application if AI in moderating harmful content online

This report is structured into six sections:

1 **Introduction**

2 **Advances in AI and its impact on telecoms and media**
An overview of the history of AI since the initial coining of the phrase in the mid-1950s, the enablers that have facilitated progress and AI's specific impact on the telecoms media and postal sectors

3 **Current approaches and challenges to online content moderation**
This section provides definitions of harmful content, explains current methods of moderating content that is defined as harmful and highlights some of the challenges of improving the capabilities of online content moderation systems

4 **Potential impact of AI on online content moderation**
This section highlights where AI can be applied within a typical content moderation workflow to augment current systems, along with some of the commercial challenges that need to be overcome when implementing a system

5 **Potential impact of AI on promoting socially-positive online engagement**
This section explains the benefits of promoting socially positive engagement to reduce the amount of harmful content online. It also outlines some common approaches to promoting socially positive engagement and where AI can play a role

6 **Policy implications**
A summary of policy implications to inform stakeholders of some of the possible areas of concern around AI-based content moderation

2 ADVANCES IN AI AND ITS IMPACT ON TELECOMS AND MEDIA

In the mid-1950s, John McCarthy coined the term Artificial Intelligence. This was based on "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".[20] Since then, AI has gone through a number of phases of hype followed by disillusionment (often described as an 'AI winter'). We are now at a stage where the necessary technological enablers and commercial applications are in place to make another AI winter unlikely.

A key technological enabler has been the increasing capability of machine learning, a subset of AI techniques. This, along with reduced cost and availability of storage and computational power, is good reason to expect AI to continue to advance and have an impact across an increasing number of sectors.

The telecoms and media sectors have both begun to adopt AI for various applications. As we transition to the next generation of mobile connectivity, 5G, it is expected that the use of AI in both of these sectors will become more pervasive with an increasing number of applications.

2.1 INTRODUCTION TO AI AND ITS HISTORY

AI refers to the capability of a machine to exhibit human-like behaviour

There are many definitions of Artificial Intelligence (AI), but a general definition is "the capability of a machine to imitate intelligent human behaviour".[21] For the purposes of this report we will use this definition to define AI as a capability rather than specific algorithms or groups of techniques, such as machine learning.

Progress in AI is still restricted to performing narrowly defined tasks

AI has multiple levels with increasing breadth and depth of capability. In each level, parallels can be drawn between the capability of the AI and human intelligence. Broadly, there are three levels:

1. **Artificial Narrow Intelligence** (ANI) performs a limited function or set of functions extremely well, often better than a human, for example image recognition or autonomously driving a vehicle.
2. **Artificial General Intelligence** (AGI) would be able to perform a multitude of tasks of interest to a similar level as a human and is indistinguishable from a person in the Turing test.[22]
3. **Artificial Super Intelligence** (ASI) refers to a system with a cognitive capability which surpasses that of humans.

These levels of AI capability are discussed in more detail in Appendix A.

Currently, humans have only developed ANI systems. Predictions for when we will first develop AGI are varied, but the general consensus is that we will not see a development of this kind for at least several decades. Artificial Intelligence within the scope of this report is, therefore, limited to ANI.

2.1.1 AI HAS BEEN THROUGH SEVERAL CYCLES OF HYPE FOLLOWED BY DISILLUSIONMENT

In recent decades AI has been through several cycles of interest and investment from research communities and society at large, subsequently followed by periods of disappointment and low interest, known as 'winters'. There have been two significant AI winters to date which are characterised by significantly reduced funding for research and public pessimism within the field. The timelines of these cycles are shown in Figure 5.

In the past decade we have seen an increasing rate of progress, sustained interest from academic research and many real commercial applications of AI. Meaningful milestones have been achieved for AI performance in various fields including voice interfaces, autonomous vehicles and cancer diagnoses. In many applications, considerable value comes from complementing human performance with AI performance.

The first AI winter was caused by optimistic predictions which could not be realised by the computational power available at the time

The first AI cycle was from 1956 to 1972, a period where research into Artificial Intelligence gained momentum as a field of study. Early researchers at this time made optimistic predictions of the future of AI which could not be realised by the computational power available at the time.

Herbert A. Simon was an economist and winner of the Nobel Prize in economics in 1978 and the Turing Award in 1975. He

predicted in 1958 that within 10 years a computer would beat the world's chess champion. Allen Newell, a computer science and cognitive psychology researcher who won the Turing Award with Simon, claimed that within the same period, a computer would be able to discover and solve important new mathematical theorems unsolved by humans. Simon went on to later predict that, by 1985, machines would be capable of doing any work a human can.

In the late 1960s Marvin Minsky, another prominent cognitive scientist and founder of MIT's Computer Science and Artificial Intelligence Laboratory, believed that AI would be solved within a generation. In 1970 he famously stated in *Life magazine*[23] that a general intelligence machine was only three to eight years away.

These researchers fuelled the media hype at the time. However, they had failed to appreciate the scale of the problems they were facing - the most significant problem was limitations in computing power. In 1976, the world's fastest computer, the CRAY-1, cost US$5 million and could process 100 million instructions per second.[24] At this level of computing power, to imitate the brain a computer would have cost US$1.6 trillion – more than the GDP of the USA.[25] By comparison, the same computational power has been available in low-end smartphones since around 2009.

Expert systems provided a short period of hope in AI research in the 1980s but the high cost of data storage triggered a second AI winter

The first AI winter lasted until the advent of expert systems in the early 1980s. These were programs that could answer questions and solve problems within a specific domain of knowledge.[26] Significantly, they proved that AI could be applied to real world problems.

By 1987 desktop computers began to overtake the capability of expert machines of the time. Cheaper, non-AI alternatives became available to consumers. Interest reduced in expert systems which required more data, faster processing power and more efficient algorithms, leading to the collapse of the expert-system industry. The AI available at the time didn't live up to the hype and interest declined, triggering the second major AI winter which lasted from 1987 to 1993.

2.1.2 ANOTHER AI WINTER IS UNLIKELY AS AI IS NOW DELIVERING COMMERCIAL VALUE ACROSS SECTORS

It is unlikely that another AI winter is coming. AI is now delivering commercial value and is supported by improvements in AI techniques, computational power and data storage capacity. Francois Chollet, a leading software engineer in machine learning and AI at Google, predicted that "There will not be a real AI winter, because AI is making real progress and delivering real value, at scale. But there will be an AGI winter, where AGI is no longer perceived to be around the corner. AGI talk is pure hype, and interest will wane."[27]

AI has grown, in part due to increasing computer power and by focusing on isolated problems. Machines now have the capabilities to classify images, understand natural language and anticipate a person's actions. These successes make AI commercially viable as there are now applications that have tangible benefits for companies. Today, AI is an integral part of many larger systems across diverse applications such as data mining, industrial robotics, speech recognition, banking software and medical diagnosis. These use cases have been enabled by advances in the underlying technologies, allowing data to be collected, stored, shared and processed in volumes that wasn't possible in the past.

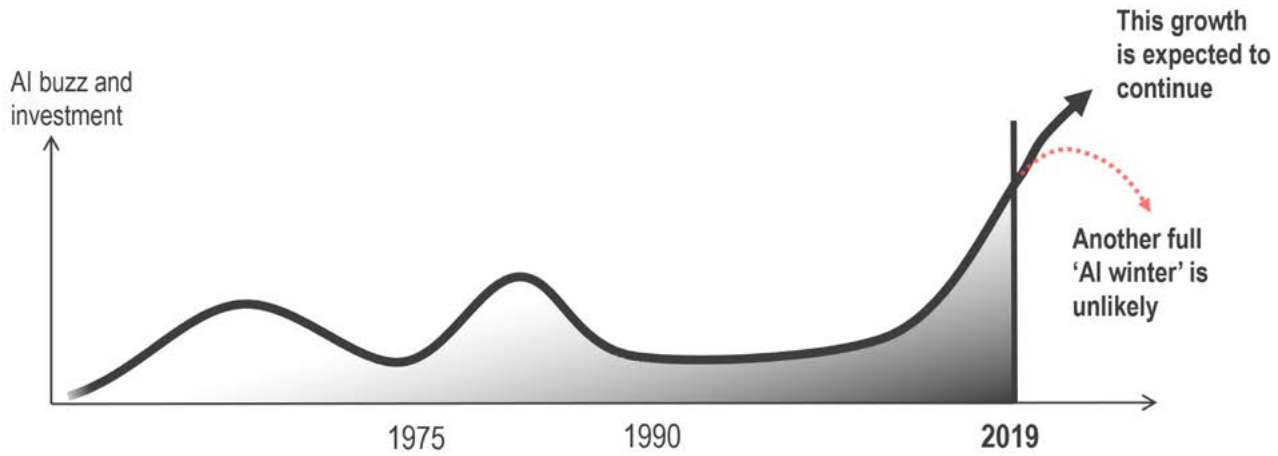


Figure 5 – AI has been through several cycles of buzz and investment followed by AI winters (**SOURCE:** Cambridge Consultants)

2.2 RECENT DEVELOPMENTS IN AI

There have been rapid advances in AI in the past decade which have led to sustained progress and increasing levels of commercialisation through real-world deployments. Figure 6 shows key milestones in the development of AI capability throughout recent years. This progress has been enabled by advances in the underlying technologies and in the algorithms used to provide AI functionality.

2.2.1 RECENT ADVANCES IN AI HAVE BEEN ENABLED BY PROGRESS IN ALGORITHMS, COMPUTATIONAL POWER AND DATA

The recent developments in AI have been enabled by advancements in three key areas, which are each described further below:

1. **Algorithms**
   Research into new algorithms and techniques has led to many new advances in AI (see section 2.2.2).

2. **Computational power**
   Computational power has increased as a result of improved silicon technology which has facilitated the development and use of advanced chipsets better suited for AI processing.

3. **Data**
   Data has become more accessible through increased connectivity, allowing data to be more easily transported, the reducing costs of storage and more collaboration in sharing data sets.

Figure 7 shows the key advances in each of these areas, illustrating the acceleration since 2010. Advances in these areas have enabled and contributed to advances in other areas, thereby enabling the acceleration of AI capabilities overall. Many of these algorithms and AI capabilities are discussed in further detail in Appendix A.

1. **AI algorithms have experienced significant development in recent years**

Researchers have pushed the boundaries of AI capabilities through the development of advanced algorithms which outperform their predecessors. Artificial Neural Networks (ANNs) and in particular Convolutional Neural Network (CNNs) have been around for many years, but recent advancements have facilitated the development of more complex networks which can achieve far superior results using innovative network architectures. For example, consider the development of Generative Adversarial Networks (GANs) which pit two networks against each other to simultaneously improve their capabilities using a feedback loop (see section 2.2.2). The application of CNNs and GANs to online content moderation is discussed in section 4.

2. **Computational power and storage have reduced in price and AI-focussed processors are becoming increasingly common**

AI techniques, and machine learning in particular, require significant computational power, especially for image and video-based tasks. Available computational power has significantly increased in the last 20 years. Figure 8 demonstrates the vast improvements in computational power, shown by the increase in transistors within a microprocessor, and the dramatic decrease in cost of processing power (measured in billions of floating point operations per second, GFLOP).

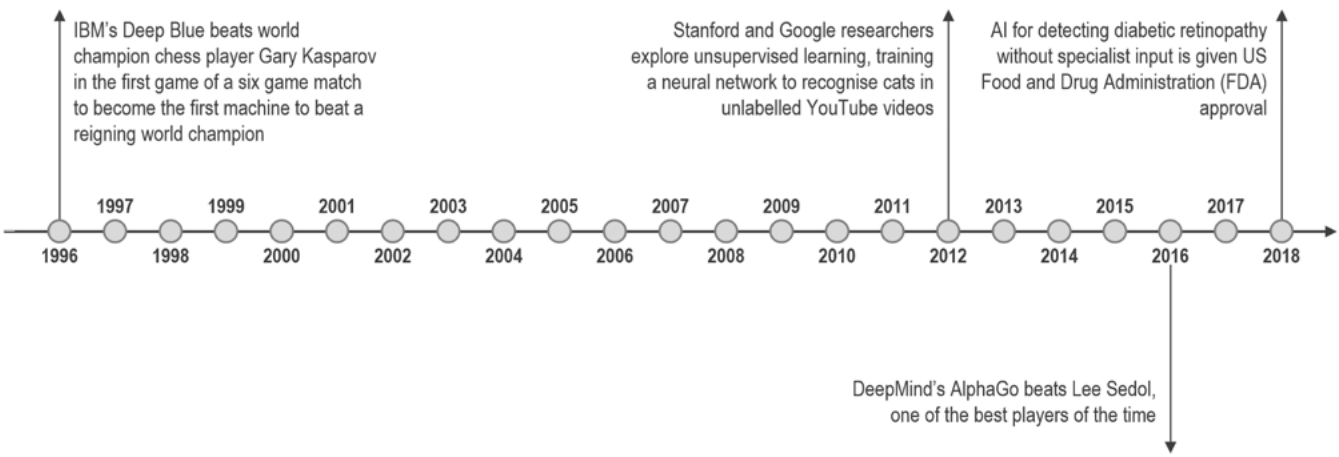
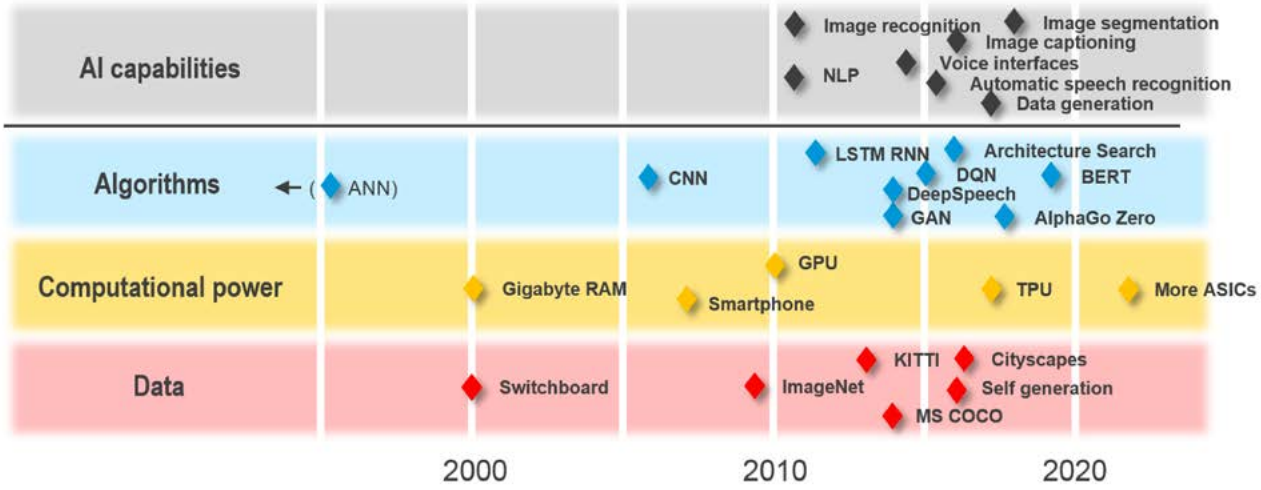


Figure 6 – Key milestones in AI capabilities have been reached at an increasing rate (**SOURCE:** Cambridge Consultants)

The silicon cost reduction, along with improved connectivity, has greatly increased the access to computational power through cloud computing. This is a growing market which is expected to reach a value of US$411 billion by 2020.[28] There has also been increased research into high-performance hardware developed with specific AI applications in mind. For example, Google's Tensor Processing Units (TPUs), are its newest AI chips which deliver 15-30 times higher performance and 30-80 times higher performance-per-watt than standard CPUs and GPUs.[29] Google claims this allows its cloud service to run state-of-the-art neural networks at scale – and therefore at a much more affordable cost, reducing the barriers to entry for a range of industry players.



ENABLER	TERM	DESCRIPTION
Algorithms	▪ ANN ▪ CNN ▪ LSTM RNN ▪ GAN ▪ DeepSpeech ▪ DQN ▪ Architecture Search ▪ AlphaGo Zero ▪ BERT	▪ Artificial Neural Network (see Appendix A.5) ▪ Convolutional Neural Network (see Appendix A.5) ▪ Long Short-Term Memory Recurrent Neural Network (see Appendix A.5) ▪ Generative Adversarial Network (see Appendix A.5) ▪ Open-source speech-to-text engine ▪ Deep Q-learning ▪ Networks to automatically identify the optimal neural network architecture during development ▪ Deep reinforcement learning AI developed by DeepMind ▪ Bidirectional Encoder Representations for Transformers
Computational power	▪ Gigabyte RAM ▪ GPU ▪ TPU ▪ ASICs	▪ Gigabyte Random Access Memory ▪ Graphical Processing Unit ▪ Tensor Processing Unit ▪ Application Specific Integrated Circuit
Data	▪ Switchboard, ImageNet, KITTI, MS COCO, Cityscapes ▪ Self-generation	▪ Readily available datasets used for AI training ▪ Synthesis of training data using Generative Adversarial Networks

Figure 7 – There has been rapid growth in the development of AI techniques in the past decade (**SOURCE:** Cambridge Consultants)

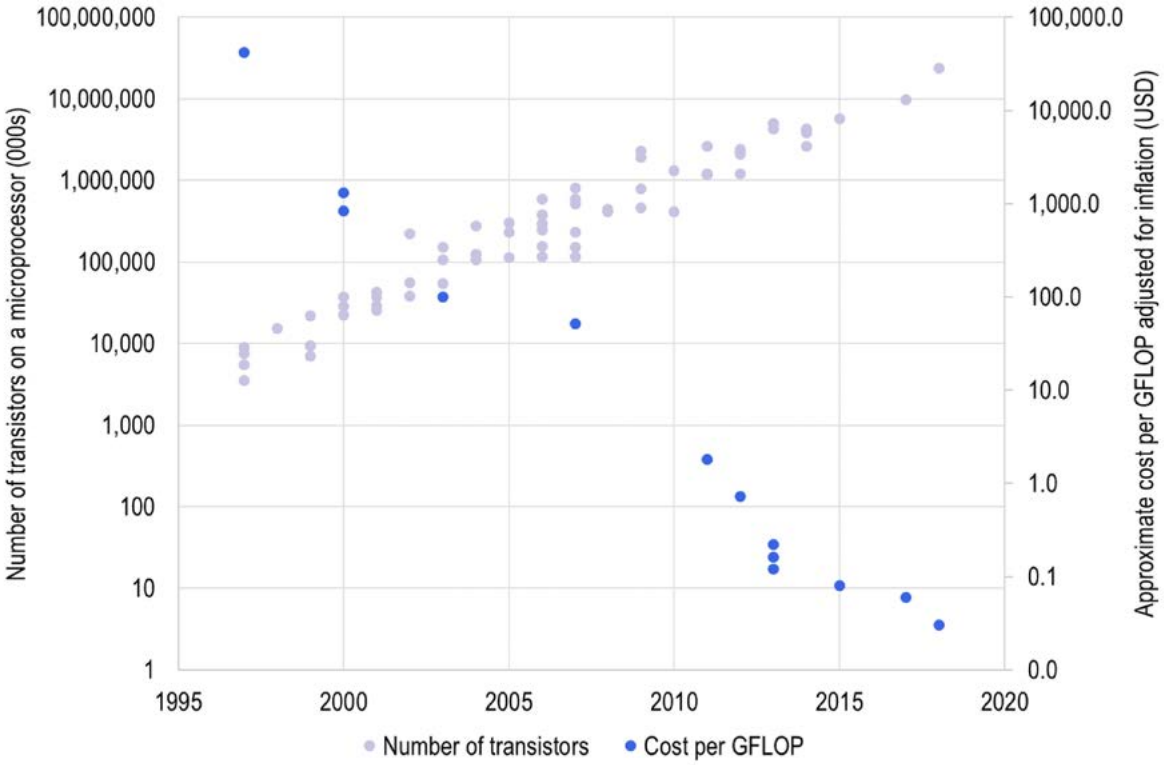


Figure 8 – The cost of computing power has reduced significantly over the last 20 years while the number of transistors on a microprocessor has greatly increased (**SOURCE:** Cambridge Consultants)

3. Access to training data has increased in the last decade, powering the new wave of AI development

With all aspects of the digital world creating vast amounts of data, training data has become much more accessible, allowing researchers, engineers and data scientists to more effectively train their AI systems. Major tech giants like Facebook and Google understand the importance of data for the advancement of AI and have made large data sets publicly available to facilitate the training and development of advanced AI algorithms. Furthermore, publicly available datasets such as ImageNet and KITTI have enabled researchers across the world to collaborate to improve their algorithms and increase their AI capabilities.

It is not possible to measure the performance of AI in comparative terms

The performance of AI has increased rapidly in recent years. However, there is no simple way of measuring this performance change because there is no single metric which encompasses all aspects of the capability of AI.

Different applications need to be measured in different ways, depending on the important features of that application. For example, accuracy is important in a system which diagnoses cancerous tumours from medical images, but speed is more important for in a control system for a drone operating in a quickly changing environment.

Even within a single application, such as online content moderation, using a standard test dataset and controlled conditions to compare the performance of systems is problematic. This is discussed in more detail in section 3.3.

2.2.2 BREAKTHROUGHS IN MACHINE LEARNING HAVE SUPPORTED RAPID ADVANCES IN AI CAPABILITIES

The latest wave of AI has mainly been driven by machine learning

Machine learning is the study of algorithms and statistical models which enable a computer system to make decisions and predict outcomes without being explicitly programmed.

Machine learning techniques have driven much of the progress in AI in recent years.

Figure 9 shows how machine learning techniques fit into the wider AI landscape, although the precise definition and relationships between different terms in the area of machine learning are often debated by the industry.

Machine learning is implemented in two very distinct phases: training and inference

Training is the process of exposing a machine to new data which it then uses to improve its performance of a task. For a machine to learn effectively, the dataset should include a wide range of examples. These will depend on the technique that is being used, and the task that the machine is being trained to complete. The training phase typically consumes a large amount of computational resource to allow the machine to analyse a large number of samples. The process of training is experimental, using trial and error, and so it is difficult to predict the resulting performance of the machine.

Inference is the process of using a trained machine to make decisions based on new data that it has not seen previously. The machine deals with a single instance at a time and so the required computational power may be small, such as a smartphone, but it will vary depending on the complexity of the task. Performance can be characterised easily and if deemed insufficient, the machine can return to the training phase.

There are three approaches to machine learning which are suited to different applications

1. **Supervised learning**
   Supervised learning is typically used to classify data. It requires a dataset where each input has a labelled output. This dataset is used by the machine to learn how to map each input to the correct output. This mapping is known as an inference function and, once adequately trained, the machine should be able to correctly label input data that it has not seen before using the inference function. For example, a computer can be shown labelled images of cats which it can use to learn features within the image that are unique to cats. It can then use the information it has learnt to recognise new, previously unseen images of cats.

2. **Unsupervised learning**
   Unsupervised learning is used to understand the structure of a dataset in the absence of labels. The machine clusters input data based on similarities but does not place input data into specific categories. The purpose of unsupervised learning is to find hidden patterns or groupings within the data. For example, a machine could group similar images without any prior knowledge of category names.

3. **Reinforcement learning**
   Reinforcement learning is a goal-orientated approach to machine learning. The machine attempts to reach its goal by learning from feedback either in the form of positive

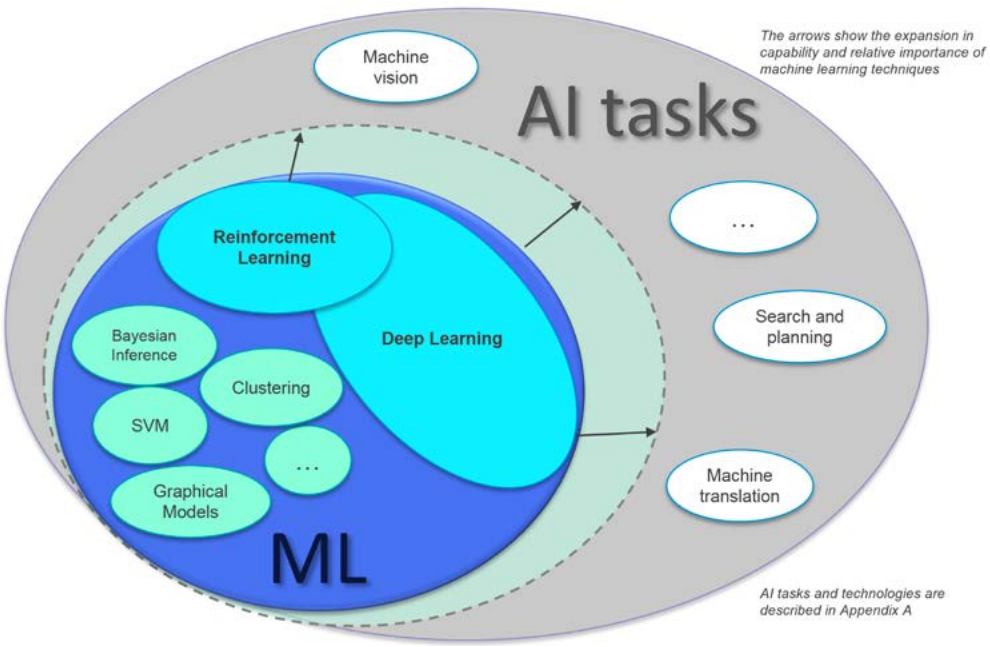


Figure 9 – Machine learning is a major area of AI and its techniques have been expanding in capability and importance relative to others (**SOURCE:** Cambridge Consultants)

or negative reinforcement. For example, reinforcement learning can be used by a machine to learn to play Pac-Man by learning to maximise its score (described further below).

In addition, semi-supervised learning methods exist which use both a small set of labelled data and a large set of unlabelled data. This is a valuable approach in applications where the cost of labelling data is high.

An important distinction between these approaches is the level of data and the training environment each technique requires.

- Supervised learning requires a comprehensive labelled dataset
- Unsupervised learning requires a large volume of input data, but the data does not need to be labelled
- Reinforcement learning does not need a dataset for training, but it does require a training environment for reinforcement where the experimentation can be carried out and there is a clearly defined metric to optimise

These approaches to learning are compared further in Appendix A.

The most significant breakthrough of machine learning in recent times is the development of deep neural networks for deep learning

Artificial neural networks are algorithms that mimic the biological structure of the brain. The majority of neural networks are 'feed forward' in which data is passed through each layer of the network, with the output of one layer being the input to the next. Deep neural networks are defined as having two or more hidden layers of neurons. The concept of neural networks has existed since the 1940s; however, the sheer amount of computing power required meant it was not economically feasible until the past decade.

Deep learning is the term to describe machine learning approaches which use large artificial neural networks. This is based on information processing patterns used in biological brains and this approach scales better than other approaches with large amounts of data. They do not require manual engineering to respond to specific features of the application they are applied to. Deep learning is an important advancement of machine learning as it enables algorithms to recognise features in data itself, allowing the analysis of complex data inputs such as human speech, images and text.

Figure 10 below illustrates how deep learning can be used for image classification. In this case it is implemented using a CNN which is a technique for layering filters in order to recognise increasing levels of complexity and abstract detail. This approach is described alongside other approaches in Appendix A.

Deep learning has contributed to improvements in understanding natural language

Deep neural networks have also made considerable contributions to common natural language processing tasks. This approach outperforms the traditional methods in named

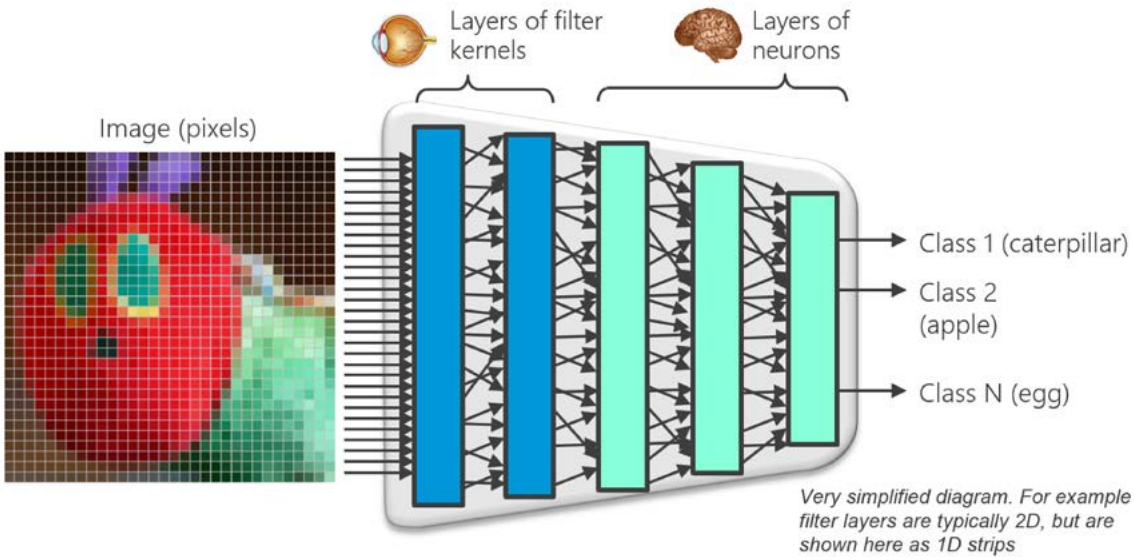


Figure 10 – A deep learning approach to image classification uses layers of neural networks (**SOURCE:** Cambridge Consultants)

entity recognition (which classifies text into predefined categories), part of speech tagging (which defines each word by its type such as noun or verb), machine translation and sentiment analysis. Natural language processing is discussed in more detail in section 4.1.1.

Reinforcement learning provides a technique to make decisions to tackle challenging problems

Another key advancement in recent years is reinforcement learning. This approach does not need large labelled datasets and can instead learn by trial and error. The power of which was demonstrated by Google's AlphaGo which beat one of the best Go players in the world, Lee Sedol in 2016. AlphaGo Zero learns by playing the game against itself, with the goal of winning the game. Go is a game with simple rules but has over 130,000 possible moves after just two moves and so is far more computationally intensive than Chess which has just 400. However, reinforcement learning techniques enabled a machine to beat a leading Go champion, which was considered decades away just a few years before it was achieved.

An example of how reinforcement learning can be applied is for learning the game of Pac-Man, which Cambridge Consultants have developed in our AI research lab.[30] In Figure 11, the right side shows the gaming environment and the left side explains the AI system's future plans, highlighted in pink.

In a similar approach to the training of AlphaGo Zero, the system played many games against itself to learn the most effective methods of playing the game. Through positive and negative reinforcement feedback, it learnt to understand which moves constituted good and bad moves and thus how to improve its gaming strategy. As the AI system learnt how to play the game through trial-and-error, it improved its game play through several stages as shown in Figure 12.

Reinforcement learning is particularly well suited to challenges where a positive outcome can be measured easily, such as scoring points in a game. To train an AI system requires a learning environment in which the agent can experiment and improve through trial and error. Despite many advantages, reinforcement learning has some limitations: whilst it may perform well for tasks such as Pac-Man which have clearly defined metrics of success, it struggles for less well-defined performance metrics on which to optimise. Furthermore, many applications cannot provide safe learning environments for reinforcement learning. For example, an autonomous vehicle could not, in practice, learn to drive experimentally on public roads due to the high cost associated with mistakes

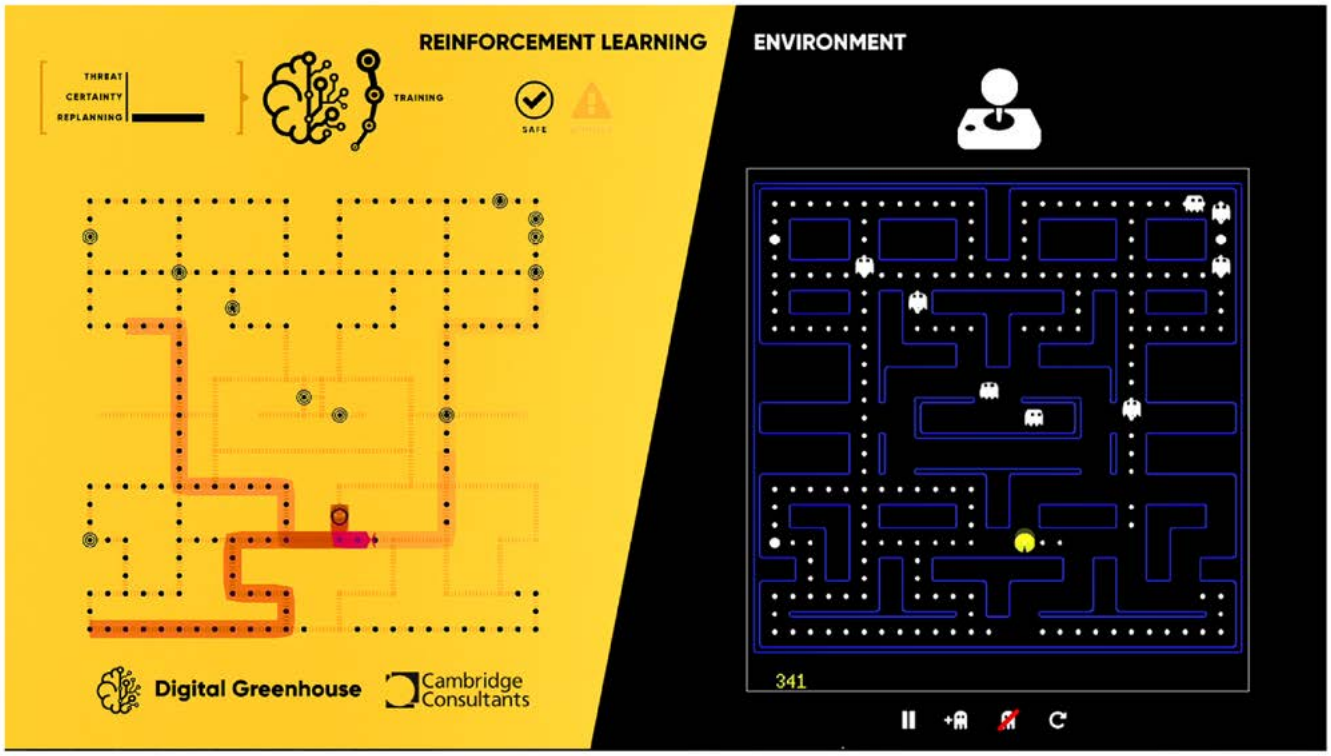


Figure 11 – An AI agent can learn to play Pac-Man using reinforcement learning (**SOURCE:** Cambridge Consultants)

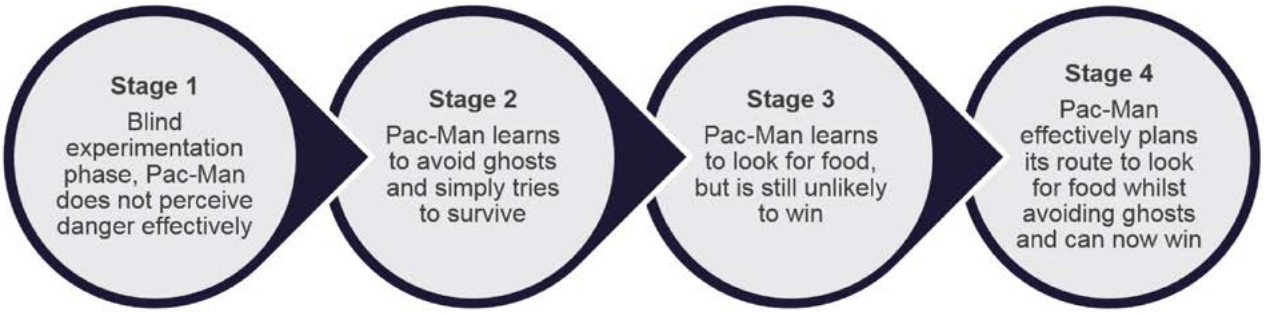


Figure 12 – As the system learns to play Pac-Man through reinforcement learning, it improves its game play through stages (**SOURCE:** Cambridge Consultants)

during training. Simulations are required not just for safety, but also to minimise costs. Reinforcement learning will likely be an essential tool in problems involving decision making, as it uses previous experience to improve the outcomes of future choices in a similar way to a human. A learning environment, or realistic simulation of one, is therefore required to train reinforcement learning models.

Reinforcement learning techniques have been applied successfully to computer vision, speech recognition, natural language processing, machine translation and many other fields to produce results comparable or even superior results to humans.

Generative adversarial networks (GANs) provide a technique to generate fake content and to detect generated fakes

The development of GANs in 2014 has enabled neural networks to 'imagine' and create new data. This is achieved through an architecture composed of two neural networks which are pitted against each other – one the generator, the other the discriminator. These networks simultaneously improve each other's capability through a feedback loop.

Facebook's Director of AI Research called adversarial training "the most interesting idea in the last ten years in ML".[31] GANs are valuable in generating new data from a smaller set of real data. This feature can be utilised to augment a set of training data, which can then be used to train a deep neural network system improving the performance of the system. GANs are also the approach used to generate Deepfakes, which are fake images or videos which appear to be genuine but are in fact synthesised by the GAN. These are discussed in more detail in section 4.2.

2.2.3 AI PROGRESS HAS BEEN SUPPORTED BY AN OPEN AND COLLABORATIVE APPROACH

The rapid pace of AI development is supported by a willingness to publish progress and support open source software initiatives

The collaborative nature of the AI industry has played a significant part in advancing AI and its capabilities. Technology companies have made remarkable progress in recent years but have also contributed to AI development through their willingness to support open source frameworks for AI development.

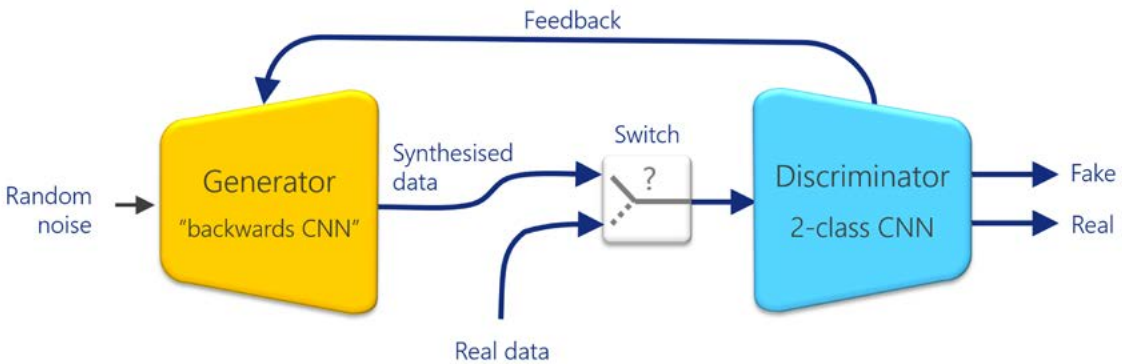


Figure 13 – GANs synthesise new content by competing with a system which learns to discriminate real content from fake content

TensorFlow, an open sourced software library developed by Google, has supported the development of AI by providing a machine learning framework which facilitates both deep learning and flexible numerical computation. Similarly, Microsoft open sourced its Cognitive Toolkit empowering developers to train deep learning algorithms by providing access to high speed, scalable computation.

There continues to be strong collaboration between academic research and industry

Whilst academia has played a fundamental role in developing the field of AI through theoretical and experimental research, collaboration with industry provides the skills and data necessary to develop real world AI applications. This continual collaboration between academia and industry has been a catalyst for AI development, with several industrial initiatives encouraging AI research in academia.

2.3 EXPECTED FUTURE OF AI

Continued development of AI is expected which will increase its impact across a number of sectors. The UK Government suggests[32] that to achieve this in the UK a holistic approach is required with consideration given to advancing the field in general by increasing access to data, developing skilled individuals and increasing research. It also highlights the importance of supporting the uptake of AI. Another government report[33] suggests that "creating an economy that harnesses artificial intelligence (AI) and big data is one of the great opportunities of our age". However, a number of barriers must be overcome in order to achieve this.

2.3.1 AI WILL CONTINUE ADVANCING RAPIDLY AND HAVE AN INCREASING IMPACT ACROSS SECTORS

Commercial applications of AI are being deployed across sectors

The capabilities of AI have been advancing rapidly in the last decade and are predicted to continue to do so. AI will be applied increasingly across all major industrial and commercial sectors. It is reported that AI will have the potential to create US$3.5 trillion to US$5.8 trillion in value annually across 19 industries.[34] The range of applications and ways in which AI will deliver value to society is very broad.

There are several areas in which AI has already become established which give an indication of the future impact it will have. Three particularly notable areas are described opposite:

VIRTUAL ASSISTANTS FOR CUSTOMER SERVICE

A major use of AI in the service sector is for virtual assistants, i.e. chatbots, for customer service. Chatbot software relies on AI as they use natural language processing (NLP) to 'understand' an input from the customer and perform a task, this is then used to generate a relevant reply. They may take in text input or speech input, which requires speech recognition and processing. Although customer service chatbots are limited in scope, for their purpose they have been very successful, and had a positive reception. When compared to human customer service in the 2016 Aspect Consumer Experience survey, the main downside of chatbots – lower accuracy in gauging the issue – was mitigated by their ease of use[35] (including involving no human interaction) and speed. Chatbots are forecast to deliver annual savings of US$11 billion globally by 2023 across retail, banking and healthcare sectors[36], through improved customer satisfaction and reduced labour and other operational costs.

AUTONOMOUS VEHICLES

AI is having a major impact in enabling the development of autonomous vehicles. AI technologies are essential for vehicles to reach level four autonomy[37], which is true self-driving while confined to an intended area. This requires the use of many forms of sensor data to detect road features and other information to control the vehicle, as well as route planning. The adaptability required for this has only become possible through using machine learning technologies. Level five autonomy requires complete adaptability to the wide range of events that could occur in real-world driving, and most experts believe we are not close to this yet.

HEALTHCARE

AI will have a major impact on healthcare and provide significant benefits to humans. In February 2019, the UK's National Health Service (NHS) updated its code of conduct[38] to make it easier for companies to work with them to develop AI and data-driven technologies to provide better health services. Accessible and personalised medical care will continue to emerge in the form of monitoring technologies and machine learning techniques that enable patient-specific predictions and treatment. AI already has a role in diagnostics, particularly in image analysis (from microscope slides to facial phenotype detection). It is currently used to aid medical specialists in their decisions and, in the future, it will have more responsibility

to make decisions based on data analysis, with specialists collecting relevant data and communicating results, whilst still performing quality assurance. AI technologies allow virtual consultations and can be used to speed up the development of new drugs in many processes due to its strength on massive datasets – often its ability to recognise hidden patterns easily.

Open-source software and collaborative approaches will support advancements and applications of AI

A number of large technology companies have released open source machine learning software. In 2015, Google released its open source framework TensorFlow, advertised as 'An open source machine learning framework for everyone'.[39] It has also recently released DeepMind Lab, a first-person 3D game platform, its purpose is to aid the development of general AI and machine learning systems.[40] Facebook has also released open source AI tools[41] as has OpenAI through its Universe software.[42] The large amount of open source resources available will increase the accessibility of AI for companies of all sizes.

The breadth of organisations which have contributed in this way – including large businesses such as Facebook and Google, and non-profit research organisations such as OpenAI and many universities – indicates that this approach is likely to continue.

The skills shortage in AI is gradually being addressed as more skilled engineers are trained

A greater number of individuals are gaining qualifications in data science with courses being offered by an increasing number of institutions. In 2017, 74% of big data courses in the US reported increased demand for places[43] and Harvard announced a new data science programme in 2017.[44] The expansion of higher education in data science will increase the access to skilled individuals for a range of companies which will allow more companies to utilise AI.

Data will continue to be collected and made available

With the growth of the internet of things (IoT), there has been an 'information explosion', enabling data-hungry deep learning AI techniques to thrive. Intel predicts that we will have 200 billion smart devices globally by 2020, up from 23 billion in 2018.[45] Additionally, for problems where real-world data is difficult or costly to collect, technologies are being developed that enable production of simulated data to supplement small datasets, or poor-quality datasets such as when quality is limited by sensors (see section 4.2).

Simulations also enable more data to be collected and AI systems to be tested safely. For example, Carla and AirSim are two simulators that have been developed as open source simulators for autonomous driving research. Not only do these make it easier produce a huge quantity of drive time for data collection, they allow for collection of data from what would otherwise be dangerous situations, as well the ability to focus on rare events.

There is also a move towards improving open source datasets for machine learning – particularly of text for NLP, after the success of ImageNet and the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in the field of computer vision. ImageNet has over 100 million images in 100,000 categories.[46] The ILSVRC ran for eight years, ending in 2017, and highlighted some major breakthroughs in computer vision – notably convolutional neural networks (CNNs) in 2012.

Computational power will become more readily available and costs will fall

Section 2.2.1 showed how computational power is more accessible than ever before due to the introduction and increasing use of cloud computing and the general reduction in cost per unit of computational power. These trends are expected to continue with worldwide public cloud service revenue expected[47] to grow at a three-year compound annual growth rate of 16.5% between 2018 and 2021. Additionally, Amazon's market leading Amazon Web Services (AWS) provided 55% of Amazon's operating profit in 2018 Q2 despite only contributing 12% to the company's net sales,[48] this suggests there is further room to reduce price in the market particularly with strong competition from providers such as Microsoft and Google.

Investment in AI-specific chips is expected to increase substantially. Venture capitalists invested US$1.5 billion in chip start-ups in 2017, almost double what was invested in 2015. UBS also predict that AI chip revenue will grow to US$35 billion by 2021, six times the amount of 2016. Over 50 start-ups, as well as tech giants such as Apple, Amazon and Google, have said they are working on AI chips.[49] These developments are likely to result in a number of competing solutions increasing accessibility and reducing price.

AI algorithms will continue to advance and new ones will emerge

The pace of development of AI has accelerated in recent years and shows no sign of slowing down. Some of the most promising techniques are briefly described here but, given the speed at which AI techniques emerge, it is very difficult to

predict where the cutting edge will be even just a couple of years from today.

Meta-learning algorithms are being developed to enable NNs to 'learn to learn'. For example, few-shot learning aims to remove the need for a long training phase and it enables learning from only a few examples. A 2018 paper[50] proposed meta-transfer learning to adapt deep NNs for few-shot learning tasks. For example, an NN trained to recognise certain animals was able to be retrained to recognise a different animal using only a few pictures, removing the need for a large dataset.

Similarly, meta-learning techniques are being used to optimise NN architectures. Neural architecture search (NAS) algorithms are important as they remove the need for highly-qualified experts to design these architectures. NAS algorithms aim to compose the best architecture for the problem. The current state-of-the-art NASNet paper[51] NAS algorithm uses a controller RNN to sample from a set of building blocks, to create an end-to-end architecture. New algorithms are being developed in this field to achieve these results faster with less processing power.[52]

The computational power required for AI is now more widely available on edge devices, such as smartphones. Mobile chips are now designed to enable AI tasks to run at low power.[53] This allows AI tasks to be run locally, enabling more powerful AI techniques and reducing the need to send data off these devices, offering greater privacy, security and reliability.

Deep reinforcement algorithms have already enabled AI to play at human-level against experts in Go and Atari video games. These techniques are useful in solving control optimisation problems, and will enable significant advances across many fields including autonomous driving and robotics.

2.3.2 THERE ARE INHIBITORS TO THE DEVELOPMENT AND ADOPTION OF AI WHICH WILL BE ADDRESSED OVER TIME

It will take time for society to build up trust in AI

Societal adoption is key for AI to have a large-scale impact. However there is currently a lack of public trust in AI. For example, people are reluctant to allow a machine to have sole control over their medical treatment – they have higher expectations of the machine's accuracy and are less likely to forgive 'mistakes'. To highlight this, consider the international media coverage that ensues following an autonomous vehicle crash, while the thousands of human related crashes that occur every day receive none. Humans are used to understanding machines and algorithms we have engineered, whereas deep learning has hidden layers and therefore can output results we cannot explain. It will take considerable time for society to become used to trusting technology such that the full benefits of AI technology can be made real.

Whilst data has become increasingly available, a lack of access to correctly labelled data can hinder AI development

For many AI techniques, especially those involving deep learning, a lot of data is needed to train systems to obtain good results, and (up to a point) the more data the better. The data should be high quality, with low noise, and will need to cover a full range of scenarios and, if it is to be used for supervised learning, to be labelled correctly. Additionally, labelling data can be a labour-intensive exercise and when data is being produced at a high rate, large datasets are difficult to check for accurate labelling.

In general, the training data needs to be specific to the task for which the system is being trained. A machine vision system which identifies cats in images must be trained with images of cats and similar animals in order to learn to identify the differences. It is therefore necessary for organisations to collect or obtain datasets which are suitable for the system's intended purpose.

The effectiveness of AI can be reduced by unconscious human bias

There is an issue in AI with human bias affecting results. This bias can arise from parameter choices in the algorithms used, and/or from the process of labelling supervised learning datasets. Even in unsupervised learning, unexpectedly biased results can come from hidden bias in the selection of data for training and validation. For example, Amazon developed an internal recruiting tool that used machine learning, which was trained on the CVs that the company received over the previous ten years. As the technology industry is heavily male dominated, the recruiting system learned to penalise resumes that contained the word "women" in the text (such as in "captain of the women's hockey team"), and the names of colleges that were all female – even though gender as a feature itself had been omitted from the data.[54] The black box nature of many AI approaches can make it difficult to analyse where the bias has stemmed from. Additionally, this is set against a backdrop of an evolving consensus by humans of what is 'fair'. It is essential to understand and address bias within an AI system. Techniques are evolving to make systems more robust to bias, and tools exist to analyse datasets to look for imbalances.

Privacy and data protection legislation may constrain the application of AI

In the UK, the Data Protection Act 2018 places limits on personal data collection and processing, through Article 22 of the EU's General Data Protection Regulation (GDPR). Personal data must be retained "no longer than is necessary" and the intent for the data must be approved and then adhered to. Even with an automated decision based on this data, consumers have the right to an explanation of the decision ("meaningful information about the logic involved, as well as … the envisaged consequences of such processing" (Article 13(2.f)). This means that unsupervised, opaque models – as deep learning models typically are – could cause issues. Technically, if the data controller has not implemented "suitable measures" to allow human intervention, the person has "the right not to be subject to a decision based solely on automated processing including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her." (Article 22(1)). This could limit the role of AI and mean that there must always be a level of human supervision in cases which significantly affect individuals.

Most AI techniques lack transparency and are therefore inherently unexplainable

Most AI techniques, including deep learning, involve hidden layers and highly complex architectures which are impossible to analyse. This lack of transparency causes the unexplainability of many AI systems which perform many uninterpretable functions and decisions before reaching their final output. The unexplainable nature of AI decision making is unsettling to many, particularly if the decisions have high impact consequences. Developing explainable AI systems is a focus of much research.[55,56] However, this gives rise to the explainability versus performance trade-off. Simpler machine learning techniques like statistical decision trees may be more explainable but are unlikely to achieve the same level of performance as more complex models like ANNs, which may include millions of neurons but are inherently less explainable.

There are still unsolved problems within the field

Some experts in the field of AI believe that progress may slow down as the capabilities of current techniques reach their limits. Geoffrey Hinton is one of the forefathers of modern deep learning, and one of the authors of the 1986 paper introducing the back-propagation algorithm that is used to optimise a wide range of deep learning algorithms. However, Hinton now believes that to progress much further in AI a completely new machine learning technique may be required. Hinton points out that CNNs are highly inefficient at learning features: neural networks require huge amounts of memory and computing power, and massive amounts of data, and still struggle with translational and rotational changes of objects.[57] While there are many techniques with the potential to progress in this area, such as capsule networks (see Appendix A), none have yet exceeded the capabilities shown by the latest CNNs on a wide scale.

2.3.3 AI DEVELOPMENTS WILL BE DRIVEN BY COMMERCIAL SOLUTIONS TO USE CASES

The commercial model for AI is not straightforward

AI is a set of techniques and capabilities which provide human-like performance. In itself, AI does not provide valuable outputs but must be applied to a suitable use case to provide a valuable solution.

The typical commercial models to support AI development are as follows:

1. **Cross-subsidised**
   Applications of AI are funded by other business models within the same organisation, such as smart speakers which are funded by retail revenue growth, adverts or product sales.

2. **Software purchase or subscription**
   The costs of specific tools developed using AI are justified by the improved business efficiency or productivity that they deliver, or the revenue which can be gained by selling the service to others who gain improved business efficiency. Examples of this are IBM Watson and Microsoft Azure services.

3. **Product purchase**
   Certain products have AI capabilities embedded into them to deliver or support valuable features. Example products include smart home systems and autonomous vehicles. In some cases these may also be cross-subsidised as described in (1) above.

The use of AI must deliver value to justify the cost of investing in it

Developing AI systems requires access to skilled staff and suitable data to train the models, which can both be expensive. This cost and the additional development complexity that implementing AI solutions entails must be sufficiently covered by the value which implementing AI brings. Early deployments of AI have therefore been focussed on mass market solutions, such as smart speakers for consumers, or on very high value functionality, such as autonomy in self-driving vehicles.

We therefore expect to continue to see most progress being made in developing AI for applications where there is a clear and valuable use case which AI is well suited to solving.

2.4 IMPACT OF AI ON TELECOMS, MEDIA AND POSTAL SERVICES

In common with all major sectors, there is a lot of interest within the telecoms, media and postal sectors about the impact AI is having and the potential value it will deliver in the future. This section outlines the main applications of AI within these sectors.

The telecoms sector is an important enabler for the use of AI across other sectors

The telecoms sector is itself supporting the growth of the use of AI by enabling more data to be collected and transmitted to the cloud. More widespread mobile coverage and faster mobile and fixed connections are a key enabler to the use of communications technologies by organisations and consumers.

New services which use AI are now possible through the availability of high speed, low latency communications. An example of this is that of digital personal assistants and smart speakers which transmit received voice commands to the cloud for processing and then respond to the user within a timeframe acceptable to a human. Similarly, many advanced applications of AI have been enabled by communications in sectors such as autonomous vehicles and medical care beyond the clinical environment.

Multi-access edge computing (MEC) is an architectural approach which provides computing capability located at the edge of the telecoms network. This allows applications and processing tasks including AI techniques to be off-loaded from the end device, but without congesting the network or increasing latency by sending data to central servers in the cloud. Typical applications for this are in video analytics and augmented reality where there is a need for large amounts of data, high computational power and low latency. MEC approach is being standardised[58] by the European Telecommunications Standards Institute (ETSI).

As the use of advanced AI services across other sectors accelerates, this in turn drives higher demand for high-performance telecoms services, thereby leading to even better communications technologies and supporting a virtuous cycle of improvements.

The operations of telecoms networks are being enhanced by AI

AI techniques are being applied in many different areas for enhancing the operational performance of telecoms networks, from the radio access network (RAN) to the core network.

In the RAN, AI is being used to optimise the use of resources across the network and in the management of radio spectrum. The International Telecommunication Union (ITU) has formed[59] a focus group for machine learning in future networks including 5G (known as FG-ML5G). It is drafting technical reports and specifications[60] and has held a number of workshops[61] with stakeholders to discuss how ML can be applied within networks.

The applications of AI within the RAN which are most noteworthy are in enabling self-organising network (SON) activities, such as load balancing and neighbour cell relation management. AI enables datasets to be analysed to forecast traffic and optimise the network dynamically for improved performance and reliability. Anomalies in network performance can be detected in real-time using machine learning approaches, meaning that errors can be addressed earlier and predictive maintenance can be undertaken before they cause more significant performance issues.

The next generation of mobile networks, 5G, relies on high levels of spectrum optimisation and re-use to enable higher data rates to be achieved. Approaches such as massive multiple-input and multiple-output (MIMO) and beamforming are used to achieve higher data rates and focus radio energy towards end users. The high level of complexity of this is prohibitive for combinatorial approaches and so machine learning techniques are being investigated which are trained on simulations to approximate optimal conditions.[62]

In the core network, AI techniques are driving improvements in orchestration, which coordinates the efficient allocation of hardware and software components for an application or service, through software defined networking (SDN). As the number and complexity of these services increase it is not possible to do this manually and so automatic approaches which use AI techniques to deal with the complexity are essential. This and other optimisations which AI enable allow telecoms operators to offer densification and increased levels of network control while moving away from the commoditised infrastructure and best-effort quality of service approach.

Media is a leading sector in the use of AI

The media sector, and in particular the digital media sector, has been leading the use of AI technologies in many areas.

The use by Amazon of algorithms to recommend products by comparing a user's purchase history with that of others is an early example of this approach. This approach has been refined by the ecommerce and online media industries to provide recommendations for users across different media formats such as text articles, video, music and user-generated content. A similar approach is used to optimise which adverts are displayed to users.

Increasing the level of engagement of users with media channels is an important element in building the user base and revenues of most media organisations. Netflix, for example, claimed[63] in 2016 that it was saving US$1billion per year through reduced customer churn, with a particular focus on engaging a user's interest within the first 60 to 90 seconds of starting to search for a movie. It has also been experimenting[64] since late 2017 with customising the preview image used to represent movies to match the interests of individual users. This has been shown to lead to increased engagement and a higher satisfaction with users, but may also be behind the negative publicity,[65] in October 2018, when some users accused Netflix of purposefully selecting images of black actors in movies to promote that movie to black users. Netflix denied that it was using the demographics of users to personalise content and that it only used viewing history to do so. It is possible that a machine learning algorithm which had not been explicitly provided with the race of users had automatically correlated other variables, such as their viewing history as suggested by Netflix.

Neural networks are being used to analyse and tag historical archives from media organisations. Google has been working with the New York Times to analyse photos and captions of 5 million photos in their archive dating back to the 1870s. Digitising this manually would be uneconomic, but the capability of machine learning to do this and the speed and scale at which it can be applied make this a feasible technique. Similarly, AI can be used to recover or enhance historical images which have been damaged or in which original data was not captured. Deep learning algorithms have been used to convert black and white video clips into colour[66] and missing parts of images can be recreated using generative adversarial networks (more information on this approach is given in section 4.2).

Recent advances in the use of AI technology to generate new content such as Deepfakes – videos, text or images that appear to be genuine but have been generated to mislead or cause harm – have increased concern in how this will impact society. This is against a background of growing concern of the impact of 'fake news' (generated by humans) on how individuals perceive information and form opinions and what this means

for democracy. The non-profit AI research company OpenAI published[67] their development of their GPT-2 model which can generate text articles which are sufficiently convincing to be credible as human output. OpenAI are so concerned about the potential misuse of this model that they have broken with their usual policy and have released only a cut-down version. The advances in this capability and the public concerns in recent years about the impact of fake content are likely to have a major impact on the media industry in the coming years.

There are many other applications of AI in the media sector and as the capabilities of AI advance then the potential applications will increase further still.

AI will deliver improvements in the postal sector

There are many different applications of AI technology which will have a growing impact on the postal sector. The increasing level of digitisation of the sector means that there are opportunities to make use of data which has been collected, and to optimise operations to reduce costs and innovate in new areas of the service. A report[68] by Accenture in 2017 examined 25 post offices around the world and claimed that US$500 million per year could be saved by investing in advanced digital technologies.

In common with the wider logistics sector, there are significant optimisations possible in routing optimisation. An affiliate of Alibaba, the China-based e-commerce and retail organisation, established an AI lab in 2016 and claims[69] that its AI technology has allowed 10% fewer vehicles to be used and led to a 30% reduction in travel distances. The level of savings possible depends on many factors including the variation in road networks and the previous level of optimisation, but the statistics indicate that AI is capable of bringing further savings to the sector as the technology progresses.

Improvements in robots, enabled by AI techniques such as machine vision, can also improve efficiencies in handling and sorting mail and parcels faster and more accurately. Advances in this area made in the industrial and retail sectors are applicable to the postal sector.

The postal sector operates a major fleet of vehicles and AI can bring significant savings by optimising fleet management. Telematics from vehicles can be used to monitor their status and performance, and this data can be used alongside maintenance records to offer improved predictive maintenance and enhancements to the performance of the vehicle fleet.

In addition, the use of vehicle platooning, in which autonomous trucks follow the one in front to allow a series of trucks to be controlled by a single driver in the lead vehicle, will provide efficiency savings. Autonomous vehicles, autonomous delivery robots and autonomous drones are all enabled by AI technologies and will offer savings over current manual delivery methods.

In common with other sectors, the use of AI in back-office support is already having a major impact in telecoms, media and post

Improvements to corporate functions and back-office support are being delivered by AI across not just the telecoms, media and post sectors but across many different sectors.

Customer support is being improved using chatbots which can deal with many of the more routine customer support enquiries. The more complex requests from customers still need to be dealt with by a human, but by screening queries and dealing with simple requests automatically, significant operational costs can be saved. For example, AT&T, a telecoms and media organisation in the US, uses an AI chatbot to handle customer interactions and CenturyLink, another large telecoms providers in the US, uses an AI assistant which correctly interprets 99% of the 30,000 customer emails received each month.[70]

AI is being used to optimise sales and marketing efforts through the analysis of organisations' customer bases. In an article[71] published by the Harvard Business Review, Epson reports that the AI assistant it uses as a sales representative has a response rate to potential customers it contacts of 51% compared to around 15% for human sales reps. As well as targeting new customers, it can be used to identify which existing customers are most at risk of leaving and enable companies to respond and reduce churn.

Financial fraud is a major source of loss to organisations across all sectors. In a 2017 survey by the Communications Fraud Control Association, telecoms operators which responded estimated[72] that they lost US$29 billion to fraud. AI is being used to combat fraud by identifying potentially fraudulent transactions and users, led by efforts in the financial services sector. In addition to identifying patterns within transactions, in some cases, fraud can be uncovered by detecting the way in which users interact with technology such as how they operate a user interface.[73]

As the capabilities of AI continue to advance rapidly, its use will become much more common across sectors, and innovations in these sectors will have a positive impact on the telecoms, media and postal sectors.

3 CURRENT APPROACHES AND CHALLENGES TO ONLINE CONTENT MODERATION

3.1 APPROACH TO ONLINE CONTENT MODERATION

In this section we consider online content moderation in terms of what harmful content is, the form it takes and the best approaches to moderate harmful content once it has been identified.

Some harmful content can be identified by analysing the content alone, but other content requires an understanding of the context around it to determine whether or not it is harmful. Analysing this context is challenging for both human and automated systems because it requires a broad understanding of societal, cultural and political factors and interpretation of the community standards or laws which determine whether it is harmful or not.

Most large online platforms have published their community standards which define the content and behaviours which are not permitted on their platforms. This provides transparency around content moderation decisions made, but for many types of harmful content a level of judgement is required to interpret these standards when applied to a specific item of content.

The variety of content types must be considered when approaching the task of moderating online content as different types may require different moderation approaches. Content is now created in many different forms which increases the difficulty of moderation. Online platforms need to understand text, images and video in order to moderate effectively, in addition to a number of evolving formats such as memes which combine various content types to create a single piece of content.

To address the challenge of moderating a variety of content types, a workflow has been created and implemented by various parties to attempt to tackle the problem of moderating content online. Pre-, post- and reactive moderation are typically used together to increase the effectiveness of online content moderation with varying degrees of human and automated moderation taking place.

3.1.1 HARMFUL CONTENT NEEDS TO BE CAREFULLY DEFINED

Harmful content must be properly defined for the effective use of a content moderation system

As internet access has become more ubiquitous, the quantity and diversity of online content has increased dramatically. To moderate user-generated content (UGC), it is important to consider each type of harmful content separately, as different moderation approaches and technical architectures may be required for each. Figure 14 groups harmful online content and behaviour into seven loose categories, together with their associated subtypes. Some content types or subtypes may fall into several categories, such as graphic sexual content which could be deemed as sexual violence (graphic) and sexual abuse (sexual). Furthermore, differing opinions will result in content types being judged differently by different user groups, highlighting the complexity of developing online content moderation policies. The characteristics of these types of harmful content have implications for the technical approach taken in designing an AI system to moderate these types on content, as described in section 4.1.3.

Some content requires an understanding of its context to identify whether it is harmful

Some harmful content can be identified by analysing the content alone, such as images which depict graphic violence or sexual activity. However, some harmful content requires an understanding of the context around it to determine whether it is sufficiently harmful to breach the standards of the site. For example, differentiating between cases of bullying between children and cases of banter between adults requires an understanding of more than a single exchange of text between the individuals. To moderate this content effectively, it is not sufficient to analyse the content independently. The context such as the history of interactions between users and any information which is known about them must also be analysed to distinguish between harmful and innocuous content. The challenges of analysing context for AI are described in section 4.1.2.

Differences in national laws mean that some types of content are illegal in some countries but not others

Due to the global nature of the internet, it is important to consider the geographical variations of illegal content, as content which is legal in one country may not be legal in another. For example, advertising the sale of firearms is legal in the US but not in the UK, and holocaust-denial is illegal in Germany but not in most other countries. This means that national differences must be considered when moderating content, which introduces complexities and requires knowledge of the location of users who post or view this material.

Policies and guidelines are used by platforms to define what is harmful

In the past, online platforms generally implemented content moderation by enforcing hidden policies which determined which content was suitable for the platform. These policies provide the framework for content moderators to define harmful content. Following significant controversy amidst online content moderation, most large online platforms have revealed these rules in the form of their community guidelines or terms of service in an attempt to provide some transparency into the rationale behind their content moderation decisions. These community guidelines aim to define harmful content such that users both understand what content is permitted as well as the reasons behind content removals. This is an important step for content moderation as the increased transparency is critical to gaining user trust in the content moderation process.

TYPE OF HARMFUL CONTENT	SUB-TYPES	EXAMPLES	DESCRIPTION			
			TEXT	IMAGE	VIDEO	AUDIO
Child abuse material	Child nudity, physical and emotional abuse, sexual abuse, exploitation	Online group chats and forums discussing and sharing child abuse material	●	●	●	○
Violence and extremism	Promoting and glorifying terrorism, incitement to violence	Terrorist propaganda videos	○	●	●	●
Hate speech and harassment	Hate speech relating to ethnicity, origin, religion, sexual orientation, gender identity, disability; harassment or stalking	Offensive/derogatory discussion, tweets, statuses, images, videos and audio	●	●	●	●
Graphic	Suicide and self-injury, violence, animal abuse	Images and videos depicting and encouraging self-harm, graphic violence and animal abuse	○	●	●	○
Sexual	Adult nudity, sexual activity, sexual abuse, sexual solicitation	Pornography and sexual abuse images and videos as well explicit adult content and conversation	●	●	●	○
Cruel and insensitive	Bullying, body shaming, insensitive behaviour	Harmful posts targeting individuals or groups using a variety of formats	●	●	●	○
Spam	Unwanted or undesirable content, including trolling	Typically unwanted comments in groups, forums or chatrooms. Also includes posts made by 'bots'	●	○	○	○

Figure 14 – An illustrative list of different types of harmful content shows that they are generally distributed in different combinations of media type (**SOURCE:** Cambridge Consultants)

KEY FINDING

Publishing transparent guidelines that govern moderation practices will help build confidence in the standards of the platform

Online sites and services offering transparency on their definition and treatment of harmful content will help to protect users and ensure that they, and other stakeholders, will gain confidence in the standards of the platform.

These sites and services should endeavour to establish the informed consent of users about the type of content which may be encountered on the platform and offer appropriate warnings. This will ensure that users are able to make informed decisions about whether the platform is appropriate and, where appropriate, understand how to flag inappropriate content.

3.1.2 VARIETY OF CONTENT FORMATS

User generated content now appears in numerous modern formats which combine several traditional formats, increasing the difficulty of moderation

To develop an effective content moderation system requires an understanding of the many different forms of online content. Each format poses significant, often distinct, challenges to the content moderation process. The most common formats are text, image and video (which combines images and audio).

However, with the rise of social media and online forums, UGC has evolved and now appears in more complex formats. Content moderation systems must not only deal with text, image and video in their traditional format but also deal with live content such as chat and video which is posted in real-time and therefore is harder to moderate automatically before users view it. Complex formats have also evolved which combine the traditional formats, such as GIFs (Graphics Interchange Format) and memes. These complex formats are described in Figure 15.

Furthermore, to gain contextual understanding, user metadata (described in section 4.1.2), where available, must also be analysed to help identify and categorise malicious users. For the context of this report, metadata includes all additional aspects of information associated with content posted online such as: available personal data, geographical location, user history, time on site, connection type, number of exchanged messages and previous moderation removals and appeals.

Live chat replicates real-world conversation and introduces complexity to online moderation

Live chat interactions occur on many online platforms. Many users are familiar with the live chat features on social networks, but an increasing amount of online interaction occurs on online chatrooms, online gaming and live streaming platforms. These platforms incorporate live chat features enabling users to communicate with hundreds or thousands of users, many of whom are unknown participants. Similarly to text moderation, live chat moderation must analyse text to determine its harmfulness. However, the real-time nature of live chat raises additional challenges compared to text moderation.

As live chat aims to replicate real conversation online, live chat flows more freely than 'posted' text content, meaning chat can escalate much more rapidly than other types of online communication. Due to the frantic nature of live chat on gaming and live streaming sites, chat messages are often short and misspelled as users aim to respond rapidly, introducing further complexity into the content moderation process. Additionally, as online communities and chat room usage evolves, so too does the language they use. It is important to consider the user base of these online chatrooms and gaming sites which may be more popular with children. Further complexities are introduced as live chat is increasingly combined with live video. For example, Discord allows nine gamers to share their screens and engage in live video chat whilst playing online.[74]

Real-time video moderation is required for content which is live streamed

Live video has gained popularity with social networks, content sharing platforms and online gaming communities. Live video moderation tools must be developed to ensure users viewing live streamed content are protected, although this can be challenging because the level of harmfulness can escalate quickly and only the previous and current elements of the content are available for consideration. To achieve moderation in real-time requires a highly optimised system which analyses both the images in the frame as well as the accompanying audio. Furthermore, true video analysis must understand the relationship between frames to infer the true sentiment and meaning behind live streamed content.

Whilst live streaming is a powerful tool for connecting users, it raises significant concerns to content moderation systems due to the demanding requirement of analysing complex content with multiple features in real time. Facebook Live has been exploited by users who wish to spread harmful content online, recently during a terrorist attack on two mosques in New Zealand[75], as well as several cases of murders, terrorist attacks and scenes of a graphic nature being live streamed

to millions of users.[76] On the other hand, the livestreamed video of US police officers shooting and killing Philando Castille was an important moment in the #BlackLivesMatter movement and Facebook initially removed the video before reinstating it (see section 3.2.1). There are benefits to society of providing livestreaming capabilities, but moderating live content appropriately is a challenge for online platforms.

Automated GIF analysis requires text and image analysis on multiple frames

GIFs combine several images to create a simple animation. Multiple frames are displayed in succession, essentially forming a short video clip. GIFs can be used to distribute short video clips online, without requiring the user to press play. GIFs are often automatically looped and as such continue displaying the 'video' as they are distributed throughout social media, forums and other content sharing sites. Again, GIFs must be carefully analysed to moderate harmful content. Simply scanning the initial frame which is displayed as the GIF is uploaded may not be sufficient as subsequent harmful frames may appear in the same file. GIFs can also be overlaid with text adding additional complexity. Automating a system to detect harmful GIFs requires text and image analysis on multiple frames, alongside contextual understanding of the content.

Memes are highly complex, contextual content which present significant challenges to AI

Memes are designed to be relatable and are typically humorous in their nature. These properties result in a content format with high virality. Memes are often imitated, replicated and shared throughout online communities often with slight variations. Memes often target user groups or cultural norms in an ironic, humorous manner but often these memes can be offensive to certain user groups.

Memes can combine seemingly innocuous text with innocent images in a malicious manner. Furthermore, innocent memes containing innocent text and images can be posted maliciously using harmful captions. One such variation of the object labelled meme format references Ray Charles' song 'Hit the Road Jack', in which images are labelled with "Jack" and "the road" to imply one person has abused the other. Whilst the text and image may be innocuous when analysed independently, the combination of text, image and reference to the song can be harmful. To moderate such content requires an understanding of the text embedded in the image, the image itself and the caption associated with the meme.

To develop a content moderation process that can understand memes and meme culture requires contextual awareness of recent events, political views and cultural beliefs. Additionally, as memes often reference other memes or other online events, they require an understanding of internet culture, which may vary significantly between user groups and evolves over time. These examples highlight the nuances of memes and the complex challenges they present to automating content moderation systems.

FORMAT	DESCRIPTION	EXAMPLES	CONSTITUENT MEDIA TYPES			
			TEXT	IMAGE	VIDEO	AUDIO
Live chat	Online text shared in real-time	Instant messenger services and online chatrooms	●	○	○	○
Live video	Video that is uploaded and distributed in real-time	Social media 'stories'	○	○	●	●
GIF	An image with multiple frames encoded into a single image file	Animated image showing a film scene	●	●	○	○
Meme	Image, GIF or video accompanied by a caption that is often shared by internet users	Object labelled image referencing a popular song or catchphrase	●	●	●	○
Deepfake	AI-synthesised images, audio and videos, and potentially text	False videos of politicians, actors and celebrities that never occurred in reality	●	●	●	●

Figure 15 – Modern online content appears in many different formats, often combining multiple constituent media types (**SOURCE:** Cambridge Consultants)

MEMES AND ONLINE MEME CULTURE

Many different variations of the 'distracted boyfriend' meme went viral in 2017. The stock photo image displays a man distracted by a woman, much to the disgust of his partner. The meme introduced the object labelling format in which objects in the image are labelled with captions. This format combining text and image can be used to convey a complex message to the audience. The girlfriend in the meme came to represent what was the advisable thing to do while the woman came to represent the more exciting option. Many thousands of variations of this meme were created by users. One Twitter user posted an image with the man labelled "youth" being distracted by the woman "capitalism" from his girlfriend "socialism".

The object labelled meme format gained internet popularity as images can be relabelled to create new permutations, often referencing previous versions. Object labelled memes are now commonplace online and on social media and have even been used by advertising agencies.[77]

Although the labelling and implicit analogies can be understood by a human relatively easily, it assumes a lot of prior knowledge and understanding which is very difficult for an AI system to replicate.

In the example above, there are multiple elements to understand in order to understand the overall meme, and each one of these is very challenging for AI:

Cultural understanding of:

- The meaning of 'avocado on toast', not just identification as a food but its stereotype as an indulgence
- The meaning of 'Millennials' and stereotypes of their motivations
- The meaning of 'a stable career and financial stability to save for a house and start a family' as desirable in the long term but requiring short-term sacrifices

Emotional intelligence to understand:

- The underlying mindset and motivations of the man
- The meaning of the facial expression of the man
- The mindset and motivations of the girlfriend
- The meaning of the facial expression of the girlfriend

Some progress in understanding memes has been made by Facebook using their machine learning system named Rosetta.[78] However, they acknowledge that there are challenges in implementing this effectively to analyse and understand complex content such as memes.

Deepfakes have the potential to be extremely harmful and are difficult to detect

Deepfakes use machine learning techniques to generate fake content which may be posted online. Deepfakes may be synthesised images, video, audio or text generated from existing data sets. This technique can be used to create computer generated versions of politicians, actors and celebrities amongst others to simulate events that never occurred in reality. Deepfakes can be a powerful yet harmful tool as they can be used to mislead audiences into believing what they see online, through altered and misleading online content. Deepfakes gained media attention in 2018 when synthesised pornographic images and videos were distributed online depicting celebrities engaging in pornographic acts.[79] Deepfakes can be used for legitimate commercial purposes, such as dubbing foreign-language films. Deep Video Portraits[80] use machine learning techniques to transfer the head pose, facial expression and eye movement of the dubbing actor on to the target actor, to accurately sync the lips and facial movements to the dubbing audio. They can be used to mislead audiences, push political agendas and create harmful online content. As machine learning techniques and accessibility to training data improves, content moderation policies must develop tools to detect such advanced content types.

3.1.3 A MULTI-PHASED WORKFLOW IS REQUIRED FOR EFFECTIVE CONTENT MODERATION

The content moderation workflow is a multi-phase process that combines automated systems and human moderators

A number of approaches are used to moderate online content:

- **Pre-moderation** describes the content moderation process in which uploaded content is moderated prior to publication. Pre-moderation is typically performed automatically by systems with minimal human input.
- **Post-moderation** describes the moderation process in which platforms proactively moderate published content. Post-moderation involves the manual review of content that cannot be definitively categorised by automated pre-moderation systems.
- **Reactive moderation** relies on community members to flag inappropriate content. Reactive moderation is typically performed by a team of human moderators.
- **Distributed moderation** is sometimes used and relies on community members rating content that does not align with community expectations which results in content being moderated without the need for dedicated moderators

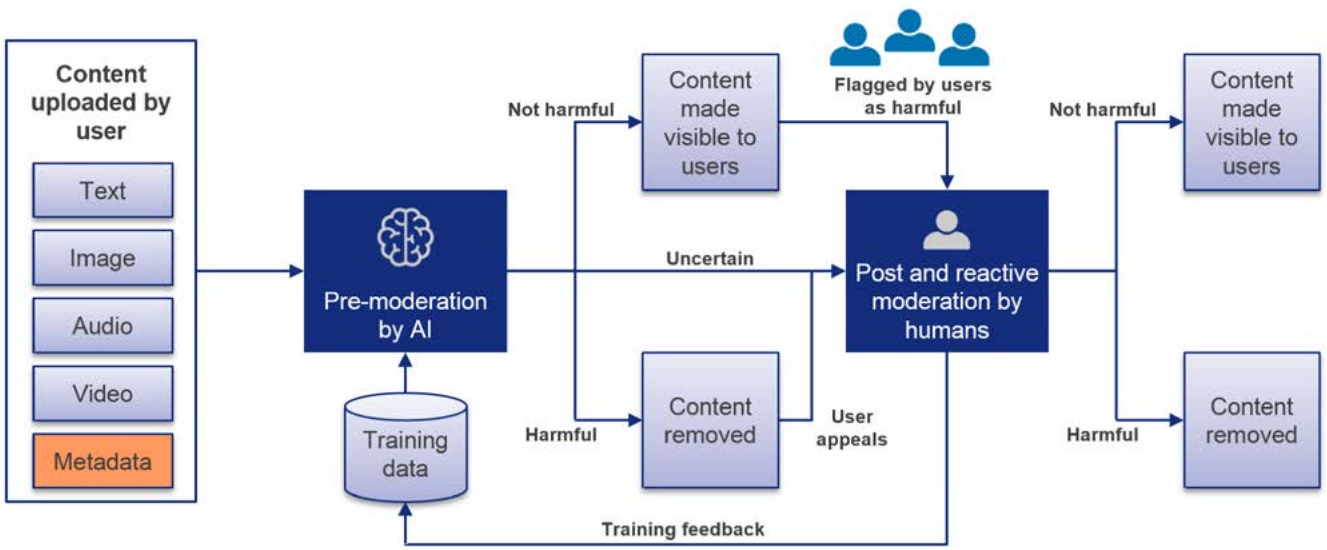


Figure 16 – The content moderation workflow combines automated systems and human moderators for pre-, post- and reactive moderation (**SOURCE:** Cambridge Consultants)

The content moderation workflow typically combines pre-, post- and reactive moderation approaches. Figure 16 demonstrates a typical content moderation workflow based on information gathered during interviews conducted with content moderation solution providers, social networks and content sharing platforms.

Pre-moderation performed by automated AI systems removes or approves uploaded content prior to publication. Automated pre-moderation is effective in removing clearly unacceptable content before it is published and is commonly used to detect illegal content such as child abuse material using hash matching algorithms such as PhotoDNA (see section 4.1.1). Pre-moderation alone is not sufficient to moderate the vast quantity of content, with complex contextual information, which is uploaded to online platforms daily. Whilst the automated pre-moderation system can automatically remove or publish content, a large percentage of content that cannot be categorised with high certainty is referred for post-moderation. Human moderators are employed to decide whether content which was flagged by the pre-moderation system is appropriate for a given platform. Post-moderation performed by a team of human moderators typically focuses on the more contextually difficult content which requires an understanding of political views, cultural beliefs, historical events and local laws across the world. The content moderation approach varies across platforms, some publish uncertain content but replicate it in a queue to be manually reviewed, whilst others will not publish uncertain content until it has been manually reviewed during post-moderation. This process depends on the size of the platform and the quantity of UGC as well as the content type and the likely reputational damage that may occur from failing to moderate effectively.

This combination allows vast quantities of content to be automatically filtered, whilst enabling the more complex content to be reviewed by a team of human moderators who better understand the nuances of online content. A significant proportion of content reviewed by human moderators occurs during reactive moderation. Reactive moderation occurs once content has been flagged or reported by the community. Many platforms integrate additional features into their flagging system to help categorise content such that it can be reviewed more effectively by the correct team.

Platforms take a variety of approaches to content moderation depending on their specific requirements

Each social network, content sharing platform and content moderation solution provider follows its own approach to content moderation. This is necessary as different sites require very different forms of moderation. A recent transparency report[81] from Google (which owns the YouTube service) shows that 81% of videos removed on YouTube between July and September 2018 were reviewed as a result of automatic flagging, the total numbers are shown in Figure 17. However, the report

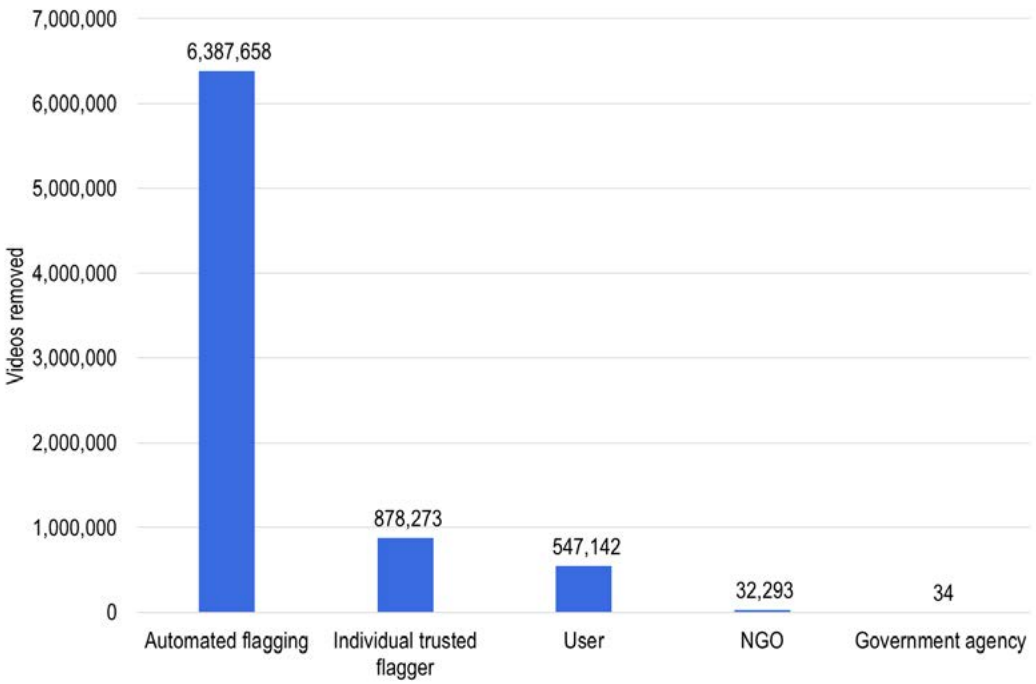


Figure 17 – The number of videos removed by source of first detection on YouTube during the third quarter of 2018 (**SOURCE:** Cambridge Consultants' representation of data from Google)

also highlights that almost 1.5 million videos required reactive moderation. It is unclear in the report how much content required human review following pre-moderation.

Automated systems remove content which is categorised as harmful above a threshold level of certainty. As demonstrated, this process removes a significant percentage of harmful content online. However, many online platforms still experience vast quantities of harmful or objectionable content on their sites.

3.2 GENERAL CHALLENGES OF ONLINE CONTENT MODERATION

The large-scale automation provided by AI means it can have a high impact on tackling the problem of online content moderation, as described in section 4.1. The amount of online user generated content is increasing rapidly, driven by improved access to online platforms through increased up-take of smartphones and faster fixed and mobile data connections. For the same reasons, access to this content is easier than ever before. Using humans to moderate content presents a significant challenge due to the sheer volume of generated content. Additionally, exposing humans to harmful online content in order to moderate it has been shown to have significant psychological effects, seriously affecting the well-being of moderators. AI can assist in this area if it is integrated well into the moderation workflow, as described in section 4.3.

However, using AI for this purpose is not without its challenges. There are a number of existing challenges to online moderation, such as ambiguity within a certain piece of content, the change of meaning within a post when context is considered and the potential for bias in the moderator. AI moderation has to contend with the same set of challenges, along with a number of AI specific challenges which are discussed in section 3.3.

A number of AI techniques can help to overcome these challenges which are discussed in section 4.

3.2.1 A CAREFUL BALANCE MUST BE FOUND BETWEEN OVER- AND UNDER-MODERATION

Online platforms must moderate to mitigate the reputational risk of exposing their users to harmful content

It is inevitable that content moderation systems will make mistakes. This can be in one of two ways: a 'false negative' when a system incorrectly identifies harmful content to be innocent, and 'false positive' when a system removes innocent content. These errors are illustrated in Figure 18 below. Note that in the context of a system attempting to identify harmful content, 'positive' means that content has been identified as harmful and 'negative' that it is not harmful.

Designing a system with a low false negative rate reduces the likelihood of exposing users to damaging content, which is a desirable feature for online platforms. To design a system with a low false negative rate typically requires increasing the threshold which determines if content is deemed appropriate or not. In doing so, the false positive rate will likely increase as the system focuses on removing harmful content. However, a system with a high false positive rate can also damage

	CLASSIFIED AS NOT HARMFUL	CLASSIFIED AS HARMFUL
CONTENT WHICH IS HARMFUL	**False negative** **Incorrect classification** Harmful content is not removed, leading to harm to viewers and damage to platform's reputation	**True positive** **Correct classification** Content correctly removed
CONTENT WHICH IS NOT HARMFUL	**True negative** **Correct classification** Content correctly remains online	**False positive** **Incorrect classification** An ineffective application of the platform's T&Cs in which content is removed when it shouldn't have been, possibly curtailing freedom of expression and damage to platform's reputation

Figure 18 – Content moderation errors can be made in two ways and these have different consequences (**SOURCE:** Cambridge Consultants)

platform reputation as users become frustrated by the removal of their content.

Low false positive and false negative rates are critical to encourage users to post and interact within their community. The reputational damage to an online platform that incorrectly moderates UGC can be huge and as such, it is important to develop a system with high accuracy – minimising both false negative and false positive rates.

It is important to note that each platform may have a different view of what is appropriate and thus each platform will have different tolerance levels or thresholds to which they develop their automated systems and train their moderators. Furthermore, each platform will have varying thresholds for different content types. For example, the reputational damage of sexual content appearing on a children's website is large and as such false negative thresholds for nudity must be extremely low. However, adult websites may be less concerned with moderating nudity and sexual activity. These thresholds must be set with consideration to the platform's audience. Where possible this should be done with the individual user in mind as well as the entire community.

Inaccurate content moderation systems can result in significant reputational damage to online platforms

In 2017, Facebook removed the image of the 16th century statue of Neptune in the Italian city of Bologna as it deemed the image to be "sexually explicit". Following the incident, the posting user, amongst others, criticised Facebook's moderation policy. Elisa Barbari who uploaded the photo stated, "How can a work of art, our very own statue of Neptune, be the object of censorship?".[82]

DRAWING THE LINE ON MODERATION IS EXTREMELY DIFFICULT

In 2016, Philando Castille was shot and killed by police officers in the USA after being pulled over in his car. By live streaming this event on Facebook, Philando Castille's fiancée sparked a huge debate about the treatment of ethnic minorities in the US and it was an important moment in the #BlackLivesMatter movement. The video was removed, possibly automatically, by Facebook prior to being reinstated. If the moderation system had been fully automated, this content may never have reached such a large audience, greatly reducing exposure and the social impact that it had.[83]

Figure 19 – An image (not the one above) of the statue of Neptune in Bologna, Italy was removed by Facebook as 'sexually explicit' (**SOURCE:** Giovanni Dall'Orto, used with permission)

3.2.2 MODERATING CONTINUOUSLY EVOLVING ONLINE CONTENT AT SCALE IS A COMPLEX TASK

The enormous scale of uploaded UGC poses significant challenges

A challenge inherent to online content moderation on a global scale, is the sheer volume of uploaded content. With several platforms boasting more than a billion users, the enormous scale of moderating UGC poses significant challenges to online platforms. A recent transparency report by Google highlighted that 7.8 million videos and 1.6 million channels were removed from YouTube in the third quarter of 2018.[84] The report also highlighted the removal of 224 million comments over the same period. Facebook revealed that it took action on 3.4 million pieces of content that contained graphic violence and 21 million pieces of content that contained nudity and sexual activity (accounting for 0.22 – 0.27% and 0.07 – 0.09% of views respectively).[85] Moderating this enormous quantity of content is a monumental task. Furthermore, with global IP traffic predicted to grow at a compound annual growth rate of 26% from 2017-2022[86], it is evident that this challenge will only increase in difficulty. Not only will automated systems play an increasingly important role in the content moderation workflow, automated systems will be critical to their effectiveness as the quantity of UGC uploaded and viewed daily cannot be solely moderated by human moderators.

KEY FINDING

Automated systems will be key for future online content moderation

Automated systems will be critical to the effectiveness of moderation systems in the future as the quantity of UGC uploaded and viewed daily continues to grow rapidly and cannot be addressed by human moderators.

"Every minute, Snapchat shares 527,760 photos, users watch 4,146,600 YouTube videos, 456,000 tweets are sent on Twitter and Instagram users post 46,740 photos" [87]

Content moderation must be contextually aware to be effective

When moderating online content on a global scale, the context of UGC is critical and must be carefully considered during the moderation process. For example, child nudity is inappropriate and would typically be removed by the majority of platforms, while a nude child being baptised may be viewed as more appropriate and may be deemed acceptable.

To demonstrate this, consider Facebook's removal of the iconic 'napalm girl' photo in Figure 20 which depicts a young nude girl running from a napalm attack during the Vietnam War. Facebook removed the photo as it breached their Community Standards stating that "while we recognise that this photo is iconic, it's difficult to create a distinction between allowing a photograph of a nude child in one instance and not others".[88] However, after widespread criticism Facebook reversed its decision and reinstated the photo stating that "in this case, we recognise the history and global importance of this image in documenting a particular moment in time". Whilst many users would agree that child nudity should be removed from online platforms, this example highlights the importance of context when moderating online content.

Context is often highly complex, requiring an understanding of cultural beliefs, political views and historical events in thousands of distinct geographical locations, each with their own diverse views defined by their varying education and environment. To address these complex issues, many online platforms employ a two-tiered approach to content moderation in which the most basic moderation is outsourced abroad where labour is much cheaper, while more complex, highly contextual content that requires greater cultural understanding is performed locally. However, the number of high-profile failures to do this properly indicates that further improvements are required to effectively moderate contextually complex content.

Figure 20 – The iconic image of a young girl fleeing a napalm bombing during the Vietnam War is of significant historical interest but also contains child nudity (**SOURCE:** Associated Press)

KEY FINDING

The ability of automated systems to understand contextual awareness is improving

AI algorithms struggle to detect harmful content which requires contextual understanding. Currently, AI algorithms are mainly deployed at the pre-moderation stage, typically identifying known harmful material using hash matching techniques (see section 4.1). Post- and reactive-moderation performed by humans requires greater cultural awareness and contextual understanding. By developing contextually aware AI algorithms, post- and reactive-moderation could be augmented with AI input, categorising and triaging flagged content.

Online content evolves rapidly over time

It is a well-known phenomenon that languages evolve over time, with new words emerging and others becoming obsolete. Further complexity is introduced in today's digital society, as more content is committed to text, making the use of symbols and code more pervasive. These can be used in a harmful or derogatory manner. One such example of this is the triple parenthesis, also known as an '(((echo)))' which is commonly used by the so called 'alt-right' and neo-Nazis to identify Jewish people in a derogatory way.[89] The symbol is commonly used on social networks to target Jews for online harassment, despite the symbol being added to the Anti-Defamation League's hate-speech database. it is evident that blacklisting keywords is not sufficient to eliminate the proliferation of spam, hate speech and harmful content online.

KEY FINDING

AI moderation algorithms must adapt to evolving content

Most current AI approaches rely on training the system with an initial dataset and then deploy the system to make decisions on new content. As harmful online content evolves, whether that be derogatory terms or phrases or a changing level of acceptability, AI algorithms must be retrained to adapt to the evolving data before being redeployed.

Moderation systems will be continually competing with users trying to subvert the system

As the rules of moderation policies and methods of moderating content are understood, people will inevitably attempt to subvert the policy or system by using slight changes to content such as text or images. This is likely to produce a continual 'arms race' between users attempting to publish content and online platforms attempting to moderate. Users are likely to use a variety of techniques to evade detection, including manually editing content or, perhaps more concerningly, using AI itself.

To demonstrate this, Cambridge Consultants has used a GAN architecture (discussed in greater detail in sections 2.2.2 and 4.2) to carry out an 'adversarial attack' to confuse an image classifier. The first image in Figure 21 is correctly identified as a police van, with a high degree of confidence (85%) by ResNet50, a well-known architecture for classifying images. However, after the adversarial attack, ResNet50 is extremely confident (98%) that the image is of a typewriter, despite the fact that the images are very similar to the human eye. This shows the potential for GANs to be used to subvert AI content moderation systems and this approach could be used to disguise harmful content. There is a growing area of research in how to devise deep learning models which are resilient to these forms of attack, such as by training the model with examples of adversarially-modified images to create a more robust system.

Figure 21 – An image classifier can be subverted by making changes which are almost imperceptible to the human eye (**SOURCE:** Cambridge Consultants)

3.2.3 AN EFFECTIVE CONTENT MODERATION SYSTEM NEEDS TO ACCOUNT FOR POTENTIAL BIAS, CULTURAL DIFFERENCES AND POSSIBLE ENCROACHMENT UPON FREEDOM OF EXPRESSION

Human bias can be introduced to AI algorithms through a number of sources

Both intentional and unconscious human bias can be introduced to AI algorithms during development. Algorithm bias can arise from the process of labelling supervised learning datasets (especially when performed manually) and under representative and biased datasets. Moreover, back-propagation and parameter choices during training can introduce or enhance bias, whether intentional or not. The 'black box' nature of many complex AI algorithms make it difficult to interpret sources of bias and thus, significant precautionary measures must be taken to ensure the effects of unconscious bias are identified and mitigated. Ensuring unbiased, representative datasets are used during training will become increasingly important to ensure potential sources of bias are mitigated.

Consider Amazon's machine learning tool for filtering and recommending the best CVs[90] as mentioned in section 2.3.2. The algorithm was trained on a set of CVs in which the number of CVs from men was disproportionally higher than from women. This resulted in an algorithm that learned to be biased towards male applicants even though the gender of the applicant was not explicitly included in the dataset. Significant bias can be unintentionally introduced to AI systems through unrepresentative training datasets. Ensuring unbiased, representative datasets are used during training is critical to the development of unbiased AI algorithms.

The recently published two-year strategy for the Centre for Data Ethics and Innovation (CDEI)[91] highlights that the UK Government is already considering the impact of data-driven services on our lives. CDEI has the remit to promote a governance environment for AI and related services and so is likely to be a key player in informing the debate around the role and application of AI-based content moderation.

KEY FINDING

Bias is not limited to human moderators

Unconscious bias in AI algorithms must be understood before automated tools can be widely deployed. AI algorithms should be trained on unbiased, representative training datasets to minimise any bias introduced into the algorithm itself.

POLICY IMPLICATION

It is important to build public confidence that any potential sources of bias in AI-based content moderation are understood and appropriate steps taken to mitigate them. This may be through auditing and calibrating datasets to understand how representative they are of the diversity of individuals in society; or by establishing a testing regime for AI-based content moderation systems.

Bias and variation between individuals leads to an inconsistent content moderation process

The content moderation process, when supported by human moderators, can introduce significant bias, resulting in an inconsistent moderation process, as illustrated in Figure 22. Content flagged for review will be manually reviewed by human moderators who may be based in different geographical locations. Each moderator will have their own views and perception of what is appropriate and as such will have their own interpretation of the guidelines by which they are governed, especially for particularly difficult, contextual grey areas. Individual sensitivities to harmful content will vary, introducing inconsistencies in the public perception of moderation. In addition, the effect of viewing lots of extreme and harmful content can desensitise human moderators and make it harder to judge the extent to which content is harmful.[92] This results in significant variation in what is deemed appropriate by content moderators due to their unconscious bias. Finally, as context evolves over time (often faster than content moderation policies can be updated) content that was once deemed innocuous may now be deemed harmful, introducing further inconsistencies to the moderation process.

This problem is heightened by the lack of feedback received by users as to why certain content has been removed whilst other, often equally harmful, content remains. This lack of transparency into the decision-making process results in an inconsistent moderation process. Increasing transparency and ensuring the basis of the decision is as explicit as possible may help remove the unpredictability and uncertainty often associated with content moderation.

AI algorithms are not subject to the same inconsistencies as human moderators but are still subject to variation

Although AI algorithms are not subject to the same sources of variation as human moderators, they still present some variation in the moderation process. AI systems may use different architectural approaches which will result in

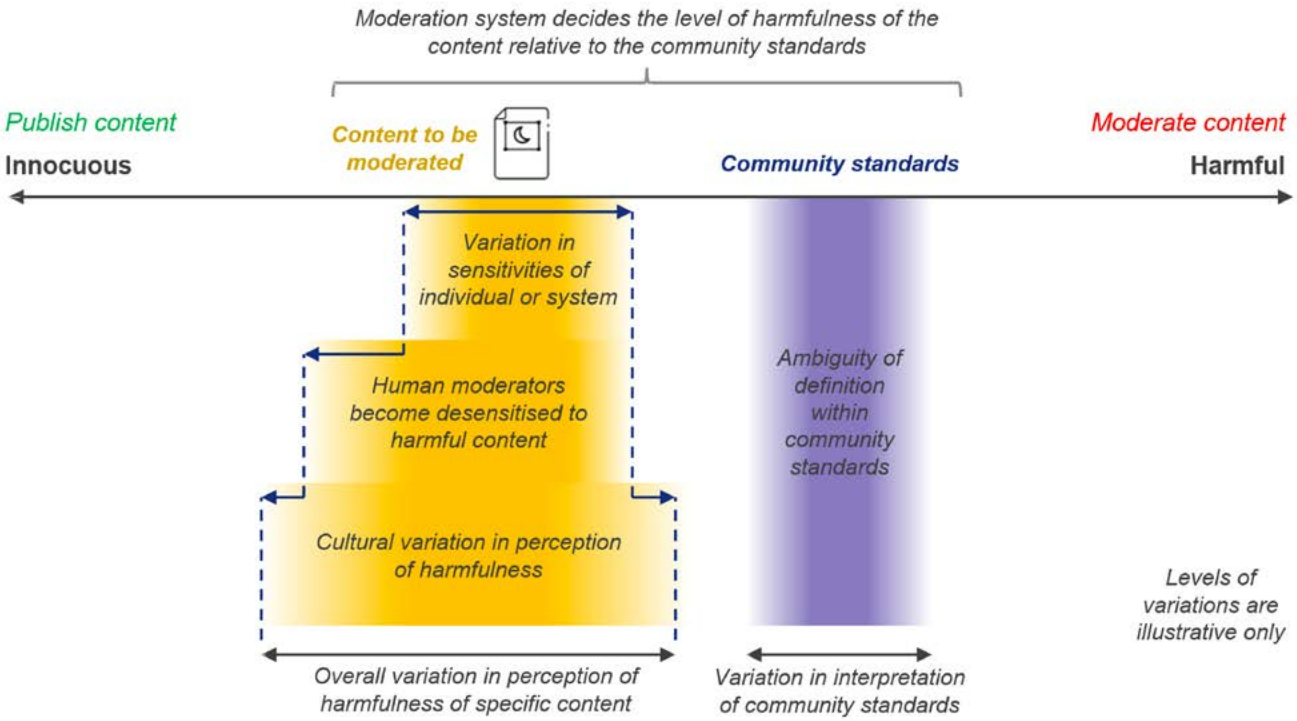


Figure 22 – Sources of variation in the content moderation process sometimes leads to inconsistent moderation outcomes (**SOURCE:** Cambridge Consultants)

different performances or use different weightings between considerations in their decision process. Even if two neural networks are trained on the same dataset they may learn to infer the function between inputs and outputs during supervised learning differently. These factors can result in inconsistent detection of harmful content across content types and platforms.

Cultural and legal differences around the world cause significant variations in what users deem appropriate

Online platforms such as Facebook, YouTube, Twitter, Instagram, Reddit and Tumblr have become content outlets for millions of people around the world. Each platform has its own version of their community standards or guidelines which govern what content is deemed appropriate.

Organisations must therefore either create and maintain separate policies for each country or apply a single policy worldwide, but each of these approaches poses significant problems.

GOOGLE'S CHALLENGE

In 2006, the Thai government requested that Google remove content on YouTube which showed the King of Thailand with feet on his head. In Thai culture this is regarded as very offensive and under Thai law it is illegal to insult the King. This was a challenge for Google because this content did not breach YouTube's community standards.

This example highlights potential problems in applying Western norms to multinational platforms.[93]

Creating effective policies for each country requires careful consideration of the cultural beliefs, political views, historical events and laws of each country. Applying these policies separately to global content is challenging and adds complexity to moderation workflows as content may be created in one country and viewed in another. Removing content in one

country because it is inappropriate in another would lead to accusations of reducing freedom of speech or of endorsing controversial regulations in another country.

On the other hand, creating a single global policy does not allow for the different cultural and legal requirements around the world. Cultural differences give rise to significant deviation in what users deem appropriate, and many users object to the one-size-fits-all approach which means that the cultural standards of one country come to dominate others.

KEY FINDING

AI algorithms must be trained on culturally relevant data

Ensuring AI algorithms are trained to the cultural norms of the countries or regions in which they will be deployed is critical to ensuring a content moderation process that accurately reflects the views of the user groups that it seeks to protect.

Moderation policies must consider internet users' rights to free speech

Many organisations in online content sharing such as Facebook, Twitter and YouTube have each developed their own community guidelines which govern the content (and to some extent) the views that they deem acceptable. However, many societies hold freedom of speech to be a core cultural value, embodied for example in the first amendment to the US constitution. It is important therefore to consider the role that platforms play in governing free speech online. People hold a wide range of views about the implications of moderating UGC and online communities' speech online, including on whether or what constitutes unacceptable restrictions on freedom of expression and on platforms' right or responsibility to curate and moderate content that connect billions of people globally.

Many would agree that content moderation is necessary, but where to draw the line is significantly more complex. Should platforms have the ability to, as some would argue, 'push' a particular set of beliefs or ideals by manipulating the content they present to their users? Others would argue that a privately-owned platform can build the community that it desires, through the content that it provides, but once a platform reaches a certain size, and becomes the 'go to' place for a certain type of content or interaction (such as YouTube for videos), it has a responsibility to society to provide a balanced approach.

KEY FINDING

The use of AI has the potential to impact freedom of speech

If the views of groups of online users are poorly or misrepresented in AI training data, AI algorithms may learn to treat them 'unfairly' or in inconsistent ways. This could potentially affect the freedom of speech of smaller online communities and minority groups.

3.2.4 MODERATION HAS A SIGNIFICANT IMPACT ON HUMAN MODERATORS AND ONLINE USER BEHAVIOUR

Moderating harmful content can cause significant psychological damage to moderators

Many organisations employ a two-tiered approach to human moderation, with basic moderation outsourced to low-cost economies, while more complex screening, which requires greater cultural familiarity is done domestically.[94] Moderators employed by content moderation subcontractors must view content that has bypassed the automated pre-moderation stage. Tasked with viewing UGC which includes the most harmful, graphic and disturbing content posted online, moderators must determine whether content breaches community guidelines.

THE HUMAN COST OF MODERATION

The psychological effects of viewing harmful content is well documented, with reports of moderators experiencing post-traumatic stress disorder (PTSD) symptoms and other mental health issues as a result of the disturbing content they are exposed to.

The Cleaners, a documentary[95] released in 2018, demonstrates the lives of these human moderators and the psychological trauma they experience as a result of their work. An article[96] by The Verge also describes the negative experience of moderators and documents how moderators have found their belief systems being altered and mental health affected by the continual exposure to extreme content.

The Technology Coalition published[97] an "Employee Resilience Guidebook for Handling Child Sex Abuse Images" which aims to reduce the damaging effects of viewing child abuse material on moderators. The guidebook states that companies should

limit the amount of time employees are exposed to child abuse material. It also states that employees should be informed of what the role entails and that sufficient programs should be in place to support employee's well-being. Clearly, content moderation serves a valuable function in protecting internet users but it is important to do as much as possible to protect human moderators from the effects of the damaging content they must review.

AI and automation may provide additional benefits to human moderators by not only reducing the amount of content that requires manual review but also by limiting the psychological effects moderating content has on moderators. Automated systems could be deployed to blur indecent images whilst still allowing moderators to make decisions, reducing the requirement for moderators to see the most disturbing elements of reported content. Furthermore, automated systems can be used to categorise and prioritise content prior to post and reactive moderation. This enables human moderators to work more effectively whilst allowing content to be distributed to the most suitable moderator for the task. Advanced categorisation can be employed to reduce the psychological damage by ensuring human moderators view a range of reported content, such that they do not end up viewing the most graphic or disturbing content for longer than necessary.

The AI techniques which can be used to reduce the harm to human moderators are described in section 4.3.

Attempts to censor information can result in the information becoming more publicised

The "Streisand effect" is the name given to the phenomenon whereby an attempt to censor information has the unintended consequence of inspiring increased interest, resulting in the information becoming more widely publicised. The term was coined after Barbara Streisand attempted to suppress photographs of her home in California from the media which resulted in greater public interest, drawing more public attention than was likely to have occurred without the attempt. Driven by human curiosity, this effect explains why attempts to over-censor information can be a challenge as people seek to find out the original information. Notable examples include the French Intelligence agency requesting the removal of a Wikipedia article about a military radio station which it deemed classified becoming the most viewed page on French Wikipedia.[98] More recently the Tide Pod challenge in which internet users ironically joked about eating laundry detergent pods resulted in some teenagers actually eating them. Attempts to censor this challenge then resulted in more widespread coverage which caused the message to reach a greater audience.

3.2.5 COMMERCIAL MODELS OF ONLINE PLATFORMS HAVE A POTENTIAL CONFLICT OF INTEREST WITH CONTENT MODERATION

Content moderation practices compete with advertising driven nature of online platforms

Most content sharing platforms optimise their design in order to increase user engagement and generate higher advertising revenues. As such, algorithms which determine trending topics and recommended videos are not neutral in pushing content but are designed to promote content that will attract users. These algorithms are designed to promote content that will have a large reach, for example it is reported that Facebook's algorithm promotes content it believes will "spark conversations and meaningful interactions between people" with posts "that generate conversation between people will show higher in News Feed".[99] Viral content captures the imagination of the community, attracting views, comments and likes which can be monetised through advertisement.

THE REACH OF DISTURBING VIDEOS

There has been a recent proliferation of disturbing videos on YouTube in which cartoons that resemble popular children's TV shows feature dark and inappropriate content not suitable for children. An account posting such content is one of the top 100 most viewed accounts in the world, with over 5 billion views.

Although this content may not be harmful to adults, there is a high risk that children view it in the belief that it is a genuine cartoon and may be traumatised by the disturbing content.[100]

Some argue that internet users are drawn to extremes, in much the same way as in the physical world. For example, many people slow down when there is a car accident on the other carriageway. This phenomenon is demonstrated online with clickbait articles which capture attention with an unusual or unbelievable headline. The problem online is that recommendation engines suggest unusual, unbelievable and obscene content as it is effective at capturing users' attention. This effect can be seen in the recent proliferation of fake news, as content is created solely for the purpose of attracting attention, regardless of whether or not the content is based on verifiable information. This causes significant problems to content moderation systems as platforms seek to encourage viral, outrageous content. Platforms moderation processes seek to balance removing harmful content with allowing outrageous, unusual content as virality generates revenue.

3.3 AI-SPECIFIC CHALLENGES OF ONLINE CONTENT MODERATION

A number of challenges specific to AI do not have to be dealt with by human moderators. These arise due to a range of factors but can generally be attributed to the higher standard of performance that is expected of AI systems compared to human moderators (an error by an AI system faces far more scrutiny than a human error, even if the human makes many more errors) and the apprehension of people generally to give up control of a system or process to AI.

Developing explainable AI systems will become increasingly important

In addition to the general challenges of online content moderation and the application of AI techniques to address these challenges, automating this process raises further challenges and complexities. Several AI learning techniques exist, from random forest and Markov models to support vector machines and neural networks. Neural networks (most notably deep learning neural networks) are inherently unexplainable. Whilst neural networks are designed to replicate the way in which the human brain learns, these networks are so complex that even the AI developer cannot understand how or why the algorithm outputs what it does. The 'black box' nature of neural network architecture which makes lots of uninterpretable decisions before its final output raises serious concerns about the application of these networks in the real world.

This difficulty is illustrated by an AI algorithm developed to identify the difference between wolves and huskies in images.[101] Whilst the algorithm was seemingly accurate at identifying which was shown in the image, investigation into the inner workings of the algorithm highlighted that it was simply analysing for snow in the background, as during training it learnt that wolves were typically photographed in snowy backgrounds.

Due to the complexity of deep learning models (which may contain millions of neurons), it is extremely difficult to develop truly transparent, explainable models. However, incorporating explainable interfaces into AI systems has been a focus of much research.[102] Neural saliency techniques[103,104] may increase the explainability of image and video moderation tools by providing importance or saliency maps during image and video analysis. These attention mechanisms can identify regions and features in the image which were deemed important by the algorithm during classification. Such techniques would help explain how an AI reached its decision by highlighting the features it analysed during decision making.

Further research into network dissection has been conducted by MIT[105,106] which aims to interpret deep learning models and their decisions. For a given image classification, an audit trail can be produced which highlights the most strongly activated path of neurons within the network, clarifying the classification process.

These techniques will become increasingly important to build trust in AI. However, explainable AI techniques are currently in their infancy and are not yet widely accessible. Alternative approaches can help build trust in these algorithms, their capabilities and their shortcomings. Ensuring representative datasets are used during training can help reduce the risk of AI algorithms learning proxies for specific features. For example, providing a dataset in which wolves were photographed in both snowy and non-snowy backgrounds would teach the algorithm to not identify the presence of wolves by simply analysing for snow.

Users have higher expectations of AI performance than human performance

A key challenge facing the automation of processes using AI is public perception. Artificially intelligent systems can analyse and process much greater volumes of data than a human and can often outperform humans at simple tasks. However, despite performing better than humans with greater accuracy and efficiency, users are much quicker to question the capabilities of automated systems that fail.

Consider the shift in public opinion of autonomous vehicles and their capabilities following a crash. Journalists and politicians have been quick to question the ability of autonomous vehicles to replace humans, despite the larger number of fatal collisions caused per mile driven by human drivers every year compared with autonomous vehicles. This heightened expectation of AI and its capabilities introduces greater complexities as humans are less tolerant to automated failure than they are to human failure. Ensuring the public's perception and expectation of artificially intelligent systems are aligned with its capabilities is critical to reducing the likely public backlash that will occur when automated systems inevitably make a mistake.

Measuring the performance of content moderation is difficult

For most forms of AI it is possible to evaluate a technique by applying it to a benchmark dataset or problem. This allows a direct comparison between different algorithms or architectures using quantitative metrics.

However, there are two problems with this approach:

1. A benchmark dataset has to be carefully selected to be representative of the real world and the type of problems the AI system is being applied to. Over time this dataset might need to evolve to keep pace with the real world, and care must be taken that the dataset is not biased such that it is unfairly better suited to one approach over another.

2. The phenomenon of overfitting may arise, which leads to an over-optimisation for the test data but the model does not generalise well to new, previously unseen data during the inference phase. In a typical AI development this quality is undesirable, and extensive work is usually undertaken to minimise overfitting. Intentional overfitting can be used to exaggerate the performance level of an AI system on a standardised dataset, but this is likely to be to the detriment of real-world performance.

A useful analogy to these challenges has been uncovered in the automotive industry where a standardised test was used to measure the level of harmful emissions from engines. Volkswagen was found to have engineered the engine software to identify test conditions and enter low emission mode, thereby providing preferential results during the tests.[107]

4 POTENTIAL IMPACT OF AI ON ONLINE CONTENT MODERATION

AI has the potential to have a significant impact on the content moderation workflow in three ways:

1. Advanced AI algorithms can be used to improve the pre-moderation stage, increasing moderation accuracy

2. AI can be implemented to synthesise training data to improve pre-moderation performance

3. AI can augment human moderators to increase their productivity and reduce the harmful effects of content moderation

These are shown in Figure 23 below within the typical content moderation workflow discussed in section 3.1.3. This section considers how AI can be used for each of these three ways. We also consider some of the commercial challenges that platforms face when implementing AI for content moderation.

4.1 AI FOR IMPROVING PRE-MODERATION CAPABILITIES

A major strength of using AI for pre-moderation of online content is its ability to process, close to real time, the huge amount of data being produced constantly by users. Every second, around 6,000 tweets are posted on Twitter[108] which would be uneconomic to review by human moderators. Moderation of data is needed in as close to real time as possible as users generally expect their content to be available as soon as they post it and viewers would be exposed to any unmoderated harmful content.

Automated pre-moderation can be split into two broad categories:

1. **Content-based moderation**
   This involves analysing the material that has been posted on its own merits without considering the wider context.

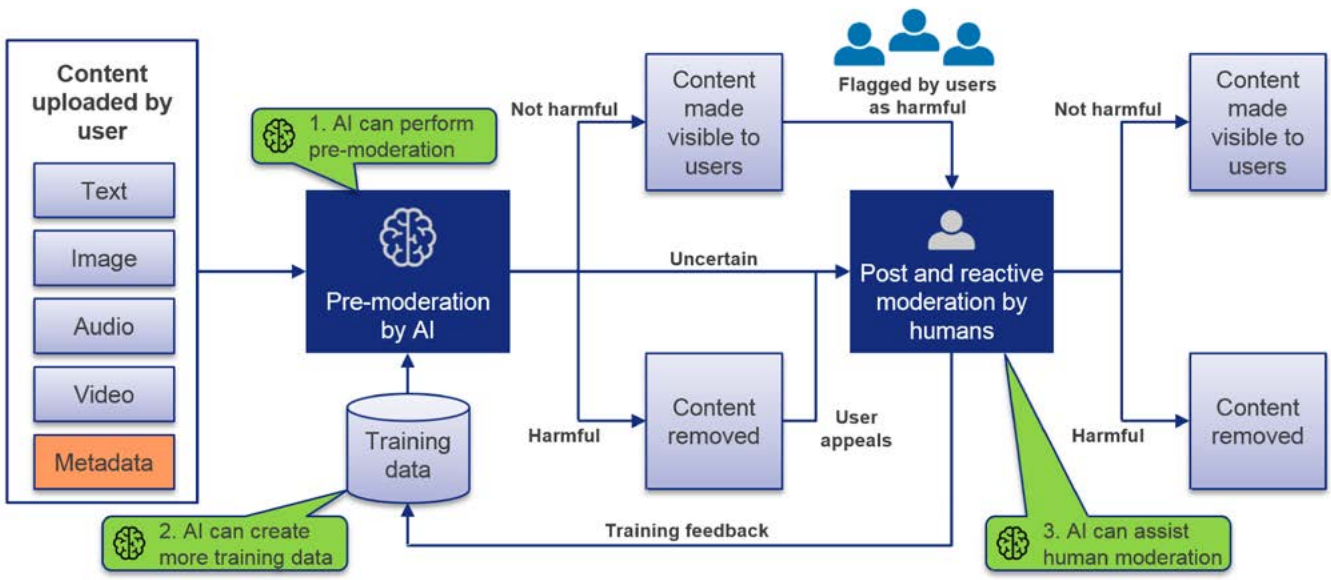


Figure 23 – There are three key ways in which AI improves the effectiveness of online content moderation

For example, this may include looking for an inappropriate object within an image, or an expletive within a piece of text. This is discussed in section 4.1.1.

2. **Context-based moderation**
   Context-based moderation also uses the context surrounding the content to analyse the harmfulness of the content. This might involve consideration of information about the user who posted the material or including the context of recent posts by other users in the same thread. Context-based moderation is a more difficult task than content-based moderation due to the much wider array of factors that need to be considered. This is discussed in section 4.1.2.

The differences between content-based and context-based moderation have a significant impact in the AI architecture which is required. This is discussed in section 4.1.3.

4.1.1 AI IMPROVES THE EFFECTIVENESS OF CONTENT-BASED MODERATION

Content based moderation techniques have seen widespread deployment by online platforms and content moderation solution providers. Techniques vary from relatively simple algorithmic tools such as hash matching and word matching, through to significantly more complex techniques such as object detection and scene understanding using AI.

Whilst simple techniques can be an effective tool for moderating online content, their capabilities are limited and as such they cannot detect all harmful online content. To increase content moderation accuracy and reduce the false negative rate requires significantly more advanced techniques, especially when the reputational damage of a false negative is great. Figure 24 illustrates an array of AI techniques applicable to content-based moderation, ranging from simple algorithmic techniques to highly complex AI approaches. The most relevant of these are discussed in this section.

Hash matching can be used to remove previously-identified harmful material

Hash matching is a computationally cheap and simple solution for removing known harmful content. Hash matching assigns a unique digital 'fingerprint' to previously detected harmful images and videos. Newly-identified harmful UGC can be automatically removed during pre-moderation if the computed hash matches a hash stored in the database of known harmful content. The algorithm used to compute the hash can provide some resistance to variations in images being used to circumvent this approach, such as mirror-images and cropped images. However, images which have been edited in more complex ways can be challenging to detect using this approach.

The Global Internet Forum to Counter Terrorism (GIFCT), founded by Facebook, Microsoft, Twitter and YouTube, aims to disrupt extremist content using a shared database of illegal content.[109] This collaborative approach allows member organisations to automatically remove previously detected extremist content using hash matching techniques. With over 100,000 stored hash values, the database has helped reduce terrorists' ability to promote extremism online.

Microsoft's PhotoDNA database, managed and supported by the National Centre for Missing and Exploited Children (NCMEC), contains previously identified child exploitation material.[110] PhotoDNA allows known child abuse material to be automatically removed during pre-moderation using computationally cheap hash matching algorithms.

Additional hash databases such as Content ID exist to regulate copyright material online, allowing copyright owners to identify and manage their content on YouTube.[111]

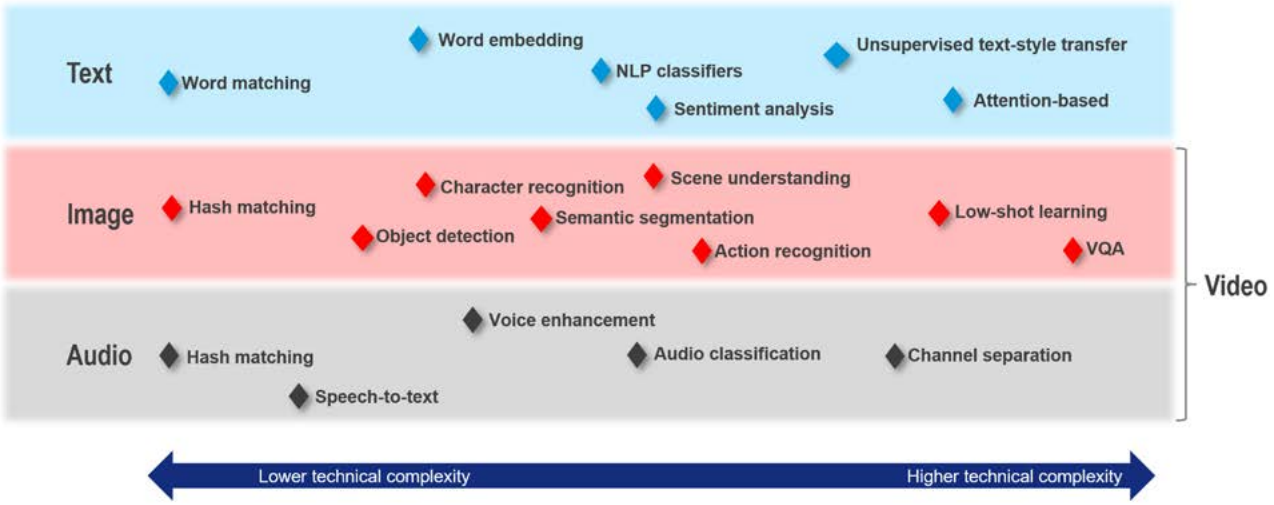
Despite its effectiveness in detecting previously identified material, hash matching cannot be used to detect previously unseen harmful content. Furthermore, extreme modifications to previously identified harmful content can circumvent hash matching algorithms, allowing previously detected content to be uploaded.

Keyword filtering is a useful tool for text moderation but has limitations

Keyword filtering is a straightforward solution to moderating text. It checks whether a block of text contains any blacklisted words or phrases stored in a database. Keyword filtering allows platforms to customise their content moderation efforts with blacklisted words or phrases that are deemed inappropriate.

Keyword filtering features similar limitations to hash matching. The list of blacklisted phrases may not be exhaustive and may lack some offensive words or phrases. Users can circumvent the system by simply modifying the spelling, and so keyword filtering requires constant maintenance of blacklists. Furthermore, keyword filtering fails to properly judge context or the intent of the text, blocking positive and harmless posts which results in a high false positive rate.

An example of the errors which keyword filtering can result in is the recent criticism YouTube faced when it deleted the accounts of several prominent YouTube content posters for allegedly uploading content that sexualises children. It was found they had posted content containing the abbreviation "CP" which is sometimes used to denote child pornography.



MEDIA TYPE	TERM	TECHNIQUE DESCRIPTION
Text	▪ Word matching	▪ Techniques to identify words by comparing them to a database of pre-defined words
	▪ Word embedding	▪ Representing vast number of unique words and sentences with a much smaller number of features
	▪ NLP classifiers	▪ Natural language processing techniques to process written text
	▪ Sentiment analysis	▪ Refers to the understanding of intent or emotion behind text
	▪ Unsupervised text-style transfer	▪ Technique to transform text into other styles or forms
	▪ Attention based	▪ Weights are given to parts of texts to represent their importance, enabling the overall meaning to be determined
Image	▪ Hash matching	▪ Technique to identify images by comparison to previously analysed and classified images within a database
	▪ Object detection	▪ Refers to the identification of specific pre-defined object classes within an image
	▪ Character recognition	▪ Machine vision techniques to identify text within images
	▪ Semantic segmentation	▪ The process of analysing an image to identify which pixels belong to which object class
	▪ Scene understanding	▪ Techniques to identify scenes within images by analysing the dimensional representation of objects
	▪ Action recognition	▪ Identifying actions of individuals/agents by observing a series of images
	▪ Low-shot learning	▪ Training computer vision models with low amounts of training data
	▪ Visual Question Answering	▪ Technique that allows AI systems to answer question about an image or text
Audio	▪ Hash matching	▪ Techniques to identify audio by comparison to previously analysed and categorised audio within a database
	▪ Speech-to-text	▪ Recognition and translation of speech into text using machines
	▪ Voice enhancement	▪ Techniques to improve voice quality
	▪ Audio classification	▪ Identifying the classes of audio sources e.g. human speech, sirens or barking
	▪ Channel separation	▪ Techniques to identify and separate audio sources for analysis

Figure 24 – There are a range of techniques applicable to moderating content in each media type (**SOURCE:** Cambridge Consultants)

In the case of these YouTube users, they were in fact enthusiasts of the augmented reality game Pokémon GO, in which CP refers to "combat points". This mistake resulted in bad publicity for YouTube's use of keyword filtering to remove harmful content automatically.[112]

NLP techniques are increasingly effective for understanding and moderating text and speech

Interpreting and understanding language within text is a challenging computational problem. Facebook's data[113] shows that only 38% of the 2.5 million pieces of hate speech removed in Q1 2018 were automatically flagged by their technology, highlighting the complexity of detecting contextually complex hate speech.

'Natural language processing' (NLP) is the term used for techniques for this and AI is contributing to advances in this field. In the moderation of online content, NLP classifiers are used to process written text and can be applied to speech by using speech-to-text techniques.

There are many NLP techniques that can be used in the moderation of text-based online content. Text based content can form the basis of harmful content such as terrorist propaganda, harassment, 'fake news' and hate speech.

Sentiment analysis is key as it classifies portions of text. This can range from simply labelling them as positive or negative, to more subtle labelling including the level of emotion. One approach to sentiment analysis is to detect relevant features in a sentence by taking the words at face-value. A technique used for this is the 'bag of words' (BoW) model, in which all words and their variants have a score which can be used to classify text as positive or negative. However, BoW approaches capture no sequential information and no syntactic content which can strongly affect the overall sentiment of the text.

'N-grams' is a similar but more successful technique, which identifies the frequency of grouped words or characters. An N-grams approach can be trained to flag up words that are misspelt or contain numbers or other non-alphanumeric characters. Although still limited, N-grams models have proven powerful in combination with other AI techniques.

'Word embeddings' is an approach which represents the vast number of unique words and sentences in a body of text, or in an entire language, with a much smaller number of features. This is achieved by grouping semantically similar words close to each other. Word embedding can also be used as a simple and efficient sentiment analysis model itself, although accuracy is often lower than more complex models. It is especially useful if there is a scarcity of large labelled datasets, as word embedding models can be trained unsupervised, then used with a small, supervised classifier.

Word embedding requires relatively little computational power and is therefore able to be trained on huge datasets, if they are available – for example Google's word2vec was trained to a high standard on 1.6 billion words, in less than a day.[114] There are several open source implementations of word embedding models such as word2vec, which is commonly-used, well-understood, efficient and accurate. It consists of a two-layer ANN using an input from a BoW or N-grams model.

Word embedding techniques are used to arrange and categorise sentences prior to their input into deep learning NLP algorithms. Deep learning models are common for text understanding tasks and can be specifically designed for sentiment analysis. RNNs and RNN-based 'long short-term memory' (LSTM) (as a later improvement) have been commonly used as they consider sequential information from the text. However, these models are computationally intensive to train.

Attention-based approaches are increasingly being used instead of recurrent neural networks for understanding language

Attention based approaches give weights to parts of the sentence representing importance, enabling the overall meaning to be determined. This approach can be used to improve RNNs by enabling them to look further back in the text, but the main benefits are that purely attention-based models are easier to train, very parallelisable, and simpler. They are also more explainable than RNNs and, at a low enough level, it is possible to understand what the model is doing.

Google Brain, a deep learning AI research team at Google, is phasing out the use of RNNs in favour of attention-based models – and their research team published a breakthrough paper in 2017 entitled "Attention Is All You Need".[115]

Detecting implications of language, such as the use of sarcasm through emojis, is difficult but successful approaches are emerging

A paper[116] was published in 2017 that had used the distant supervision of emoticons contained in tweets, to perform emoji prediction, sentiment detection, and sarcasm detection. The emoticons were used to classify the tweets without needing annotations by humans in a dataset containing over 1.2 billion relevant tweets. This vast amount of data meant that a richly expressive deep learning model could be trained with a low risk of overfitting. The model first uses word embedding, and

then deep learning methods (an attention-based LSTM). The model outperformed the state-of-the-art on all three tasks it was benchmarked on – one version of the model achieved an average of 75% accuracy on a sarcasm detection dataset.

In order to compare the performance of this model with that of human users, almost 2,500 test tweets were labelled on 'polarity' by 10 users in the US, and an average human label was calculated for each tweet. The model's results were then compared to those of the humans', and it agreed with their average on 82.4% of the tweets. The average agreement level for each person, to other people, was only 76.1%. This indicates the model was better at giving the average human-sentiment rating than a single person.

EMOJIS CAN SIGNIFICANTLY ALTER THE MEANING OF TEXT

The following three examples show how the addition of a single emoji significantly alters the tone of the message. This highlights the difficulty of text moderation where three quite different sentiments are being expressed using the exact same words but adjusted with the use of an emoji.

I love how you don't text back 💔
I love how you don't text back 😡
I love how you don't text back 😜

There are knowledge bases available to assist AI approaches to understand the meaning of complex or evolving words and phrases

Increasingly, there are also resources being developed to be used to help AI models detect relevant features, often for social good and aimed at combatting hate speech. Wikipedia has publicly available lists specialised towards subtypes of hate speech, such as ethnic slurs and slang terms for members of the lesbian, gay, bisexual, and transgender (LGBT) community. There are knowledge bases too – for example the augmented knowledge base for bullying, BullySpace, developed by MIT researchers, which contains over 200 assertions based on gender and sexuality stereotypes.[117] However, models trained on some of these can then become confined, for example to a specific subtype of hate speech.

AI can provide ways of translating and understanding content in different languages

The international nature of online content means that it is necessary to moderate content in many different languages. This applies to both human and automated content moderation systems. The limitation for deep learning AI is that there is insufficient labelled online data in languages other than English for the system to learn. However, there is promising work on language text analysis that requires almost no translated training data – including at Facebook as mentioned previously.[118] This advance in machine translation could be used to massively reduce the difficulty of the task for human-based moderation too.

Object detection, scene understanding and semantic segmentation techniques are required for moderating complex image and video content

Object detection, semantic segmentation and scene understanding are machine vision technologies that enable machines to detect the presence of objects and scenes within images and videos. Object detection and semantic segmentation use image processing techniques to identify regions of an image or video and associate this with a pre-defined class. Unlike object detection, scene understanding analyses objects in context with respect to the 3D structure of the scene by considering the global structure of the image or video.

Object detection can be achieved using two distinct approaches: classical machine vision techniques and deep learning using deep neural networks. Classical machine vision techniques use feature extraction techniques such as gradient orientation (the directional change in colours and their intensities within an image) to generate a Histogram of Oriented Gradients (HOG) before using Support Vector Machines (SVMs) for object classification. Deep learning approaches however, can achieve end-to-end object detection without the requirement of specifically defining object features, typically using image processing architectures like CNNs.

Object detection and semantic segmentation are important techniques for content moderation as they can be trained to detect and identify harmful objects and their location within an image. Additionally, memes can be detected using optical character recognition to identify and transcribe text within images.

These techniques are capable of inferring object information from images in a computationally efficient way. They can be used to identify weapons, body parts, faces and text within images and are an essential building block of content moderation architectures, together with scene understanding. Furthermore, the latest AI techniques for object detection can now outperform humans, as demonstrated at the ImageNet image classification event in 2015.[119]

For the best performance these techniques must be trained with varied input images to represent the breadth of images that a trained system would likely encounter during inference. However, collating this training data can be time consuming and expensive, especially when objects must be manually classified prior to AI training.

Recurrent neural networks can enable advanced video understanding

Recurrent neural networks (RNNs) allow the understanding of frames in a video relative to the preceding frames and are an important development in AI for true video understanding. In an RNN architecture, the output of the previous training step becomes the input into the next training step, together with the new image to be analysed. This sequential understanding is useful for any time series data like audio and video which require an understanding of the sequence itself, in addition to the individual frames or notes.

Unlike feed forward networks, RNNs utilise a feedback loop to ingest their own outputs as inputs. The RNN architecture effectively gives memory to the ANN, in which the sequential information is preserved in the network's hidden nodes, enabling RNNs to infer dependencies between events separated by many moments. For example, RNNs can be used to classify music genres, as they are able to understand the importance of the dependencies between notes as well the notes themselves.

RNNs are an important approach for content moderation as they can enable advanced video understanding, as videos and their meanings to a human are much more complex than the sum of their independent frames. Furthermore, RNNs facilitate action recognition, in which the actions of humans can be detected and analysed through time.

4.1.2 AI MODERATION TECHNIQUES CAN CONSIDER CONTEXT BUT THIS IS COMPLEX AND CHALLENGING

The context can be used in addition to the content itself to indicate how harmful that content may be. In many cases the context provides an essential element in determining the intent of the content, although this depends on the category of harmful content and the purposes and features of the website it is posted on.

Analysing and interpreting context correctly makes online content moderation a complex task. Context can cover a broad range of variables such as requiring historical or geographical knowledge (can be country specific, or specific to even smaller areas), it can depend on gender, sexuality, age, religion, race and language. There has been limited research in machine learning techniques to analyse text which takes into account the broader context. Some work has been done on more complex, but still very narrow, situations including work on cyberbullying aimed at lesbian, gay, bisexual, and transgender (LGBT) stereotypes.[120]

Context can also change over time, such as referencing the latest news stories or using the latest slang, or even the most recent post by another user. Some research[121] has been done on training a context-based ANN model to classify tweets by considering the tweets in the history stream and tweets using the same hashtags. Whilst this did improve upon state-of the art sentiment classification, the approach was confined to Twitter and may not work as effectively on other platforms.

Metadata must be analysed to detect contextually complex content types

Metadata is the data which gives information about the content and which may be considered as part of the moderation decision. The term 'metadata' has specific meanings in some technical protocols but in this report we use it to mean data which is available and relevant to the moderation process. The detection of contextually difficult content types such as hate speech and cruel and insensitive material requires advanced analysis of not only the content itself, but also the associated metadata. The metadata is critical to understanding and detecting content that is only harmful when considered in the context of the exchange.

Examples of metadata which may be used in content moderation decisions are shown in Figure 25 opposite.

The content and context metadata are illustrated in Figure 26 opposite for an example post to a social media site. It displays the content itself (the image and text) but also the additional information that accompanies the post such as the username, the posting time, the group in which it is posted, the number of likes and the number of comments. This metadata can be analysed to understand the context of the content, which can often be crucial to determining its harmfulness.

However, metadata is not limited to that which is publicly visible in an online post. Metadata can include additional information associated with the user such as their IP address, their time on the platform and their previous content history (together with previous content moderation decisions) as well as their connection type and other user identifiable information. It is common for online platforms to collect data about users to enable them to provide more relevant content and to maximise their value for the business model of the platform provider.

CONTENT	CONTEXT (METADATA)		
Posted content: ▪ Text (including any hashtags) ▪ Image ▪ Audio ▪ Video ▪ Title	**Context of the post:** ▪ History and previous posts of the thread ▪ Date and time ▪ Number of 'likes' from others (and their history and interests)	**Context of the user identity:** ▪ IP address and geographic location ▪ Device and browser used ▪ Username ▪ Real identity ▪ Age ▪ Registration information such as email address or mobile number ▪ Other information held about the user (such as friends lists, followers, interests, etc.)	**Context of the user's history:** ▪ Length of time registered on the site ▪ Amount of activity on the site ▪ Previous posts which have been moderated as harmful

Figure 25 – There are many different sources of information which relate to the content and context of UGC

Figure 26 – An example post to a social media site illustrates the content and context metadata available

Analysing the UGC and the associated metadata is important in many cases to detect harmful content, as a moderation system must be able to differentiate between ironic and sarcastic posts and comments, jokes between friends and varying degrees of harmfulness that depend on factors such as the geographic location and the user community in which it was shared.

The metadata available varies between platforms, based on their purpose and how users typically interact with content and with each other. Platforms may therefore use metadata as an input to their content moderation tools which are bespoke to their platform. However, it is difficult for platform-agnostic content moderation tools (see section 4.4.4 to make the most of the metadata which is specific to each platform. This limits the ability of content moderation services provided by third parties to fully account for the context possible in some feature-rich online platforms.

URL matching can identify malicious links to harmful content

It is common on many online platforms for users to post Uniform Resource Locator (URL, meaning a link to an internet resource) to content on another webpage or website. In many cases these links are shortened using a URL shortening service such as bitly.com to reduce the number of characters used, but this also makes it harder for the user to see where clicking the link will take them. In some cases this can be used to direct users to harmful content.

URL matching is a technique used to detect known spam and malware links. These malware links can be detected with a direct domain lookup or a blacklist check within a trusted database. There are many organisations that maintain these databases and provide this service. For example, VirusTotal allows a user to check a link against more than 60 databases including Google, Trustwave and CyberCrime. This allows users to check if a link should be trusted before entering any private information or potentially downloading any malware. However, conflicting verdicts of the safety of a website are sometimes found from different databases, requiring expert knowledge to assess the best course of action.

Algorithmic techniques can be used to identify malicious users

In practice, the majority of internet users are responsible individuals and most harmful content is generated by a small proportion of users. Therefore, identifying malicious users or their characteristics can be very valuable in detecting harmful content.

A graph-based technique can be used to identify malicious users by examining users' social networks.[122] For example, to find 'trolls' (users who consistently post negative and harmful content), users can be ranked according to a given metric (such as the number of dislikes on their UGC) and the worst ranking users can be identified as potential malicious users. A graph can be generated to show the relationships between different users and the positive/negative interactions taking place. This list of suspicious users can then be further reduced by repeating the same ranking process using a different metric. For example, by ranking the number of negative interactions a suspicious user has with other non-malicious accounts. This process is repeated, with each iteration removing more 'innocent' users.

This can identify a list of bad or suspicious users. However, this requires labelling of each interaction, whether by 'likes', 'upvotes', or manually. This also assumes that these labelled interactions are accurate and that users do not maliciously target other users by marking their interactions as negative.

AI techniques can be used to identify harmful content that may be too ambiguous to be identified without context

The techniques applicable to the content (see section 4.1.1) can generally be applied to the metadata. The outputs of which would be a confidence that that piece of data is suspicious or harmful. These scores would then be combined to generate an overall level of harm for the user and the content itself considering its context. For example, sentiment analysis could be used on all their previous content. Any sudden changes of sentiment could indicate harmful behaviour when used in conjunction with analysing the content itself.

Behavioural analytics can also be used. Understanding who the user is, the differences between other users and identifying any anomalous behaviour may help identify harmful content. If the content is borderline, this additional information from examining the context could help classify the content. A verified Twitter account with a million followers is much more likely to be who they claim to be and less likely to behave maliciously than a newly created account with no followers.

Research has shown that metadata is valuable in identifying harmful content but has limitations

Metadata can be especially useful for AI moderation of spam content or for finding fake accounts. AI can be trained to detect suspicious activity, such as an account reaching out to many more other accounts than usual, lots of seemingly automated activity, or the account having a different geographic source to the geographic location claimed by the account.[123]

Some types of harmful content can see a peak in occurrences triggered by certain events. For example, peaks in hate speech online have been shown to occur in the first few hours after a related event has occurred such as a terrorist attack.[124]

In some cases, metadata has been found to be only very weakly indicative of sentiment characterisation of specific content. However, even as a weak indicator it can be used together with other stronger indicators to improve performance by using 'ensemble' techniques to combine data sources.[125]

Research on other types of metadata being used (including number of previous posts by the user or number of replies to the specific post) has had conflicting results. This is because of the dependence of metadata on the source of the data (which can differ a lot, for instance from celebrity accounts to personal ones). As with many techniques reliant on data quality, which includes representativeness and diversity, machine learning will only be able to handle a dataset well if the model has been designed to deal with potentially poor-quality training datasets.

KEY FINDING

Human input will be required to augment AI systems for the foreseeable future

When analysing the complex contextual content posted online, it is evident that the content moderation process will require the input of humans for at least the foreseeable future, particularly for the moderation of highly contextual, nuanced content. Whilst automated systems may be effective at identifying and categorising obviously harmful and inappropriate content, it currently struggles to understand more complex content. Often such content which requires an understanding of the political views, cultural beliefs and historical events of a geographical location to fully understand the context. Whilst automated systems will improve with the development of advanced machine learning capabilities, human moderation teams will inevitably be required both now and in the foreseeable future to interpret contextually difficult content.

Regulation can inhibit the collection of user data for behavioural analysis and collecting representative datasets for training

Whilst targeting the user rather than the content itself provides benefits, there are several limitations to behavioural analytics based on collected user data. Data protection regulations constrain the collection and analysis of user data. This reduces the ability of platforms to profile and categorise malicious users and therefore impacts their ability to reduce the amount of harmful content online.

Data protection regulations can also limit the ability of moderation services to collect representative data for training AI systems. If training datasets are not sufficiently representative then there is a risk that AI systems learn and perpetuate any underlying bias in the data (see section 3.2.3). However, collecting data about users to ensure that data is gathered from a representative selection of society, such as gender, race, sexual orientation or political views, is restricted.

DATA PROTECTION REGULATIONS RESTRICT ORGANISATIONS' ABILITY TO IDENTIFY MALICIOUS USERS

A leading AI-based content moderation service provider told us that "[data protection regulations] are currently a barrier to protecting children online", as regulations restrict the ability to collect and use data sets that contain personally identifiable information. This inhibits the tracking and profiling of users engaged in sharing harmful content such as child abuse material across platforms and reduces the ability to share user data for AI training and development.

4.1.3 DIFFERENT AI ARCHITECTURES ARE REQUIRED FOR IDENTIFYING DIFFERENT CATEGORIES OF HARMFUL CONTENT

Content-based and context-based moderation approaches are each valuable to implementing content moderation using AI techniques. The balance is important between these approaches and the specific architecture required to detect harmful content depends on the category of harmful content being considered.

To illustrate this, we consider the AI architectures required to detect harmful content in two different categories: child abuse material and bullying material. Each category of harmful content will require a different overall architecture based on the specific characteristics of that content and the strengths and weaknesses of the techniques available.

Identifying child abuse material requires consideration of the content but in general context is less relevant

To detect child abuse material requires a dedicated architecture that combines content-based techniques as illustrated in Figure 27. This is based on our considerations of the techniques available and the characteristics which define child abuse material. Other technical architectures are possible and will evolve as AI techniques advance.

Hash matching should be employed to automatically remove previously detected and labelled child abuse material. To detect new, previously unseen child abuse material requires more complex screening to analyse and understand the scene.

Object detection will isolate objects and individuals for further analysis which include mood detection, age estimation and the detection of nudity and body parts. Mood detection can interpret the emotion of the identified individuals while age estimation can identify the presence of a child. This feature analysis should be combined with the detection of nudity and body parts to produce a comprehensive understanding of the image or video in question. Each subsystem will output unique confidence values outlining the detection of a particular feature. These confidence values must be integrated by considering the relative weighting of each feature and these weightings should be developed and tuned to optimise the accuracy of the system.

Concatenation of all feature outputs will generate a confidence level which describes how likely the content is to be child abuse material. This confidence level can be compared with pre-determined threshold values to trigger subsequent actions such as the automated removal, automated flagging for human review or the automated approval, allowing content to be posted on the site.

Identifying bullying content requires full consideration of the context of the user interactions as well as the content itself

Content that requires contextual understanding is much more difficult for AI to moderate. Detecting bullying is an example that requires contextual understanding and often requires image and text analysis together with metadata analysis to truly infer the sentiment behind an online interaction. Metadata such as the number of exchanged messages, connection type and age can be indicative of bullying, and this must be combined with the broader user metadata for analysis prior to classification. For example, the text 'You look like this' can range from being complimentary to offensive depending on the accompanying image. Neither the text nor the image may be offensive by themselves but once combined the content can take on an entirely different meaning. This is further complicated by the fact that the exact same caption and image combination could be complimentary or offensive depending on the parties involved. For example, a man being told that he looks like a male celebrity may be a compliment, but if the exact same content was sent to a woman it may be meant offensively, even though the content is identical. This is where metadata can play a role, by considering past behaviour of individuals, the interactions between them and the context behind the interaction.

To identify bullying in which an internet user posts an innocuous image with fairly innocent text requires an ensemble approach in which multiple networks analyse for

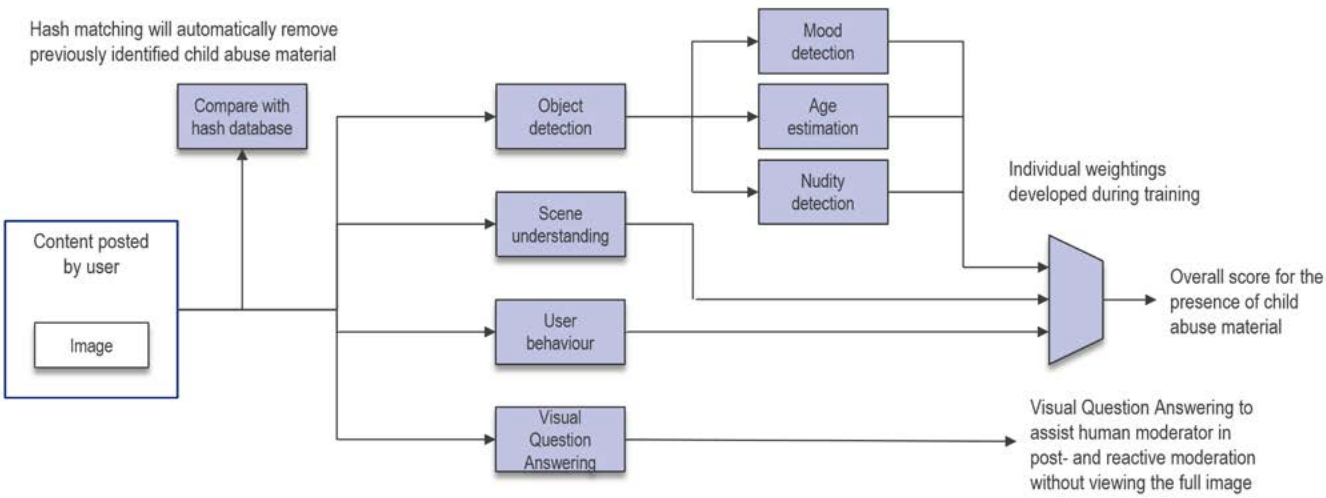


Figure 27 – An example technical architecture for detecting child abuse material illustrates the different techniques which are applicable (**SOURCE:** Cambridge Consultants)

specific features before being concatenated to analyse the overall content and context for classification. This approach is demonstrated in Figure 28.

CNNs enable machine vision techniques like object detection and scene understanding, which are critical for image analysis. These techniques can detect the presence of harmful images and locate the harmful object of interest within an image. However, object detection and scene understanding alone are insufficient to detect some examples of bullying, as the standalone image is unlikely to be harmful by itself. These techniques, therefore, must be combined with NLP techniques to identify and categorise the associated text, to detect the presence of any terms and phrases that are harmful. In addition, NLP must detect the sentiment behind the text and must be capable of identifying nuances in language which can completely alter the meaning of the interaction.

The associated metadata should be analysed using an ANN to identify patterns in the data which may be indicative of bullying. For example, the metadata could demonstrate that one user is much older than the other, or that previous interactions between the two were moderated for cruel and insensitive behaviour.

Once all networks have provided outputs of their classification of the associated inputs, an ensemble approach should then combine the outputs of individual networks for further analysis. A fully connected neural network should analyse the unique combination of content and context together to infer whether the interaction constitutes bullying. This network would require a diverse training data set of previous identified bullying examples combining positive text and positive images, negative text and negative images and all other combinations together with metadata that indicates harmful interactions.

Additionally, the network should be trained on metadata associated with bullying examples such that it learns to recognise patterns in data which are indicative of bullying. This can only be achieved with a deep learning architecture as deep learning models can extract feature data by themselves (to identify what constitutes bullying from training data) as it would be impractical to classify and categorise all combinations of text and image which may constitute bullying manually. By careful design of the AI it is possible to architect a solution that collects the diverse data types and draws insight from the singular fused information. It therefore maintains the essential contextual information that cannot be derived from the parallel analysis of individual data inputs alone.

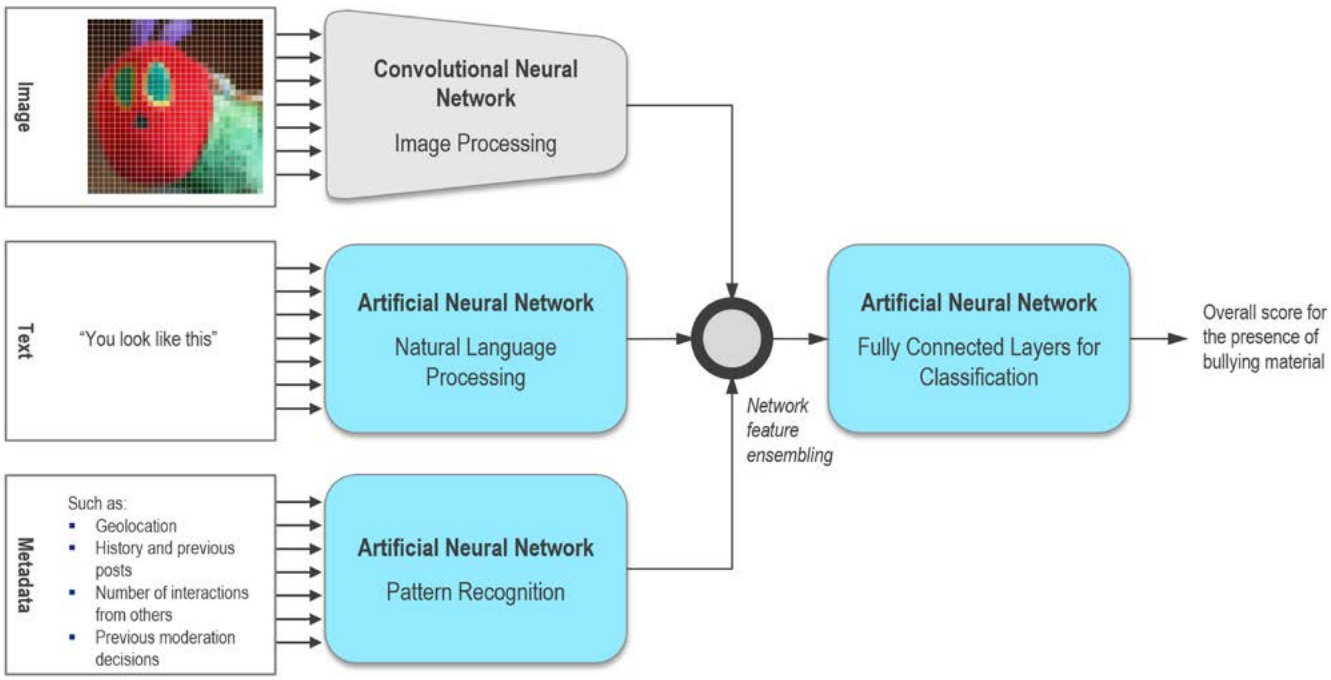


Figure 28 – An example technical architecture for detecting bullying material, which requires contextual understanding, shows how each input needs to be analysed separately and then analysed together (**SOURCE:** Cambridge Consultants)

4.2 AI FOR SYNTHESISING DATA TO AUGMENT TRAINING DATA SETS FOR MODERATION SYSTEMS

GANs can be used to augment training data by generating new and original data

GANs are a neural network architecture that can generate realistic examples from a smaller dataset. Having more varied realistic examples is essential for many AI applications, especially where the availability of training datasets is limited.

Nvidia have demonstrated the use of a GAN to generate brand new celebrity-like faces based on the CelebA-HQ dataset[126] shown in Figure 29. They used a GAN to 'imagine' new faces and create entirely new data that is still representative of the original dataset.

Plausible passages of text can also be created using GANs, which are not only grammatically correct but can provide a consistent style and tone of language. The non-profit AI research company OpenAI has announced[127] the development of their GPT-2 model for generating text. Alongside the many positive uses of this technology, OpenAI also notes the potential for it to be used maliciously such as creating misleading news articles, impersonating others or automating the production of abusive content. They have therefore decided not to release the full code of their model.

Even more contextually difficult data such as memes can also be created. Researchers as Stanford University have produced an app[128] than can take any image and produce a humorous and relevant caption.[129] The system can be conditioned on not only an image but also a user-defined label relating to the meme template.

GANs can also be used to apply style transfer to create additional data

Style transfer is technique to allow images taken in certain conditions to be transformed to others. For example, changing an image from day to night, winter to summer, or from black and white to full colour. This is achieved using GANs trained on data which has been transformed in this way such that it can then suggest a transformed version of an original image presented to it.

An example system using this technique[130] is shown in Figure 30 in which multiple different transformations have been convincingly applied to images.

Figure 29 – Realistic and convincing celebrity faces have been artificially created using GANs (**SOURCE:** Nvidia, used with permission)

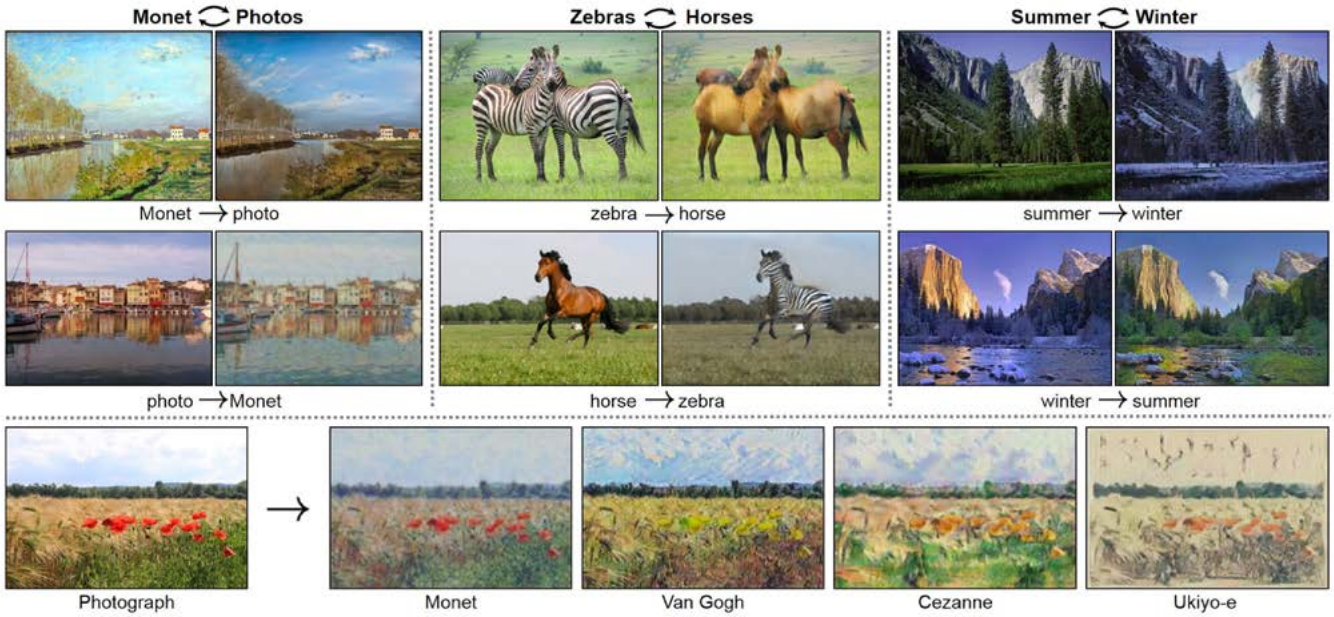


Figure 30 – Style transfer allows new images to be created in different styles (**SOURCE:** Berkeley AI Research Laboratory, UC Berkeley, used with permission)

GANs can also repair incomplete or damaged data

GANs can be used to fill a gap or replace a missing piece of data. This could be useful when data that is being used at the training or inference stages is obscured or damaged. To demonstrate this, Cambridge Consultants have created a system that is able to repair images and see beyond human vision. The transformation, shown in Figure 31, is able to reconstruct the image of an obscured plane using a GAN architecture that had been trained on images of aircraft (but it had not seen this specific image during its training phase). This approach could be used in a content moderation system to detect the content of images that have been deliberately obfuscated or damaged in an attempt to circumvent the content moderation process. This could be particularly useful for improving hash matching techniques, which aim to cross reference UGC with previously detected harmful content.

Generating more data out of a limited dataset is a valuable approach

The precise nature of what GANs are generating is subject to discussion by researchers. Most of the evaluation of the output of generators trained with the GAN framework is qualitative: authors normally list higher sample quality as one of the advantages of their method over other methods. It is hard to systematically verify the output and the samples it produces because verification might depend on the existence of perceptually meaningful features.

Figure 31 – A GAN architecture can be used to recover an image which has been damaged beyond recognition to a human eye (**SOURCE:** Cambridge Consultants)

There is a general consensus in the field that the following statements hold true:

- Generative models can approximate the distribution of real data and generate fake data that has some variety and resembles real data
- The real data has useful properties that can be extracted computationally

In addition to using GANs, a dataset can be augmented using traditional techniques. For example, given an image, a rotation, shear transform, or change in hue or colour could be applied. For text, word-swapping, syntax-tree manipulation work, sentence shuffling, or synonym insertion can change the text, without distorting its meaning. For audio, normalisation or noise can be added. These minor changes to training data allow the transformed data to be regarded as distinct data to train the network, in effect, providing a larger dataset.

GANs can reduce bias in datasets by generating content of an under-represented minority

The problem of bias in datasets is highlighted in section 3.2.3. GANs can be used to reduce the level of bias within a dataset by generating data for under-represented minorities to produce a representative training dataset. One paper[131] highlights how GANs can be 'used to approximate the true data distribution and generate data for the minority class of imbalanced datasets'. The paper shows the success of these algorithms when compared with other standard methods.

It is clear to see how this approach could be utilised for online content moderation. If a dataset contains a lack of 'edge cases', then it could be biased towards 'obviously harmful' data with a lack of consideration to data that is harmful but is more difficult to detect. Reducing this bias, by creating new data, could improve the effectiveness of a moderation system for a particular category of content.

There are limitations to the capabilities of AI in generating new training data

Biases in the AI system are introduced mainly through training data, whether that be unrepresentative or incorrectly labelled datasets. AI works best when provided with a large quantity of varied data, which has been labelled accurately. AI models will often take advantage of whatever information will improve accuracy on the dataset, especially any biases which exist in the data and this therefore amplifies any bias which is in the training data. Identifying and minimising bias in AI systems is essential for humans to trust AI systems. Methods to remove or at least reduce bias amplification exist, but they require careful application during the development of the AI system.

FACEBOOK'S DEEPTEXT

In 2016 Facebook introduced DeepText, their deep learning text understanding model, for character level and word level learning.[132] DeepText uses both CNNs and RNNs, along with word embeddings, to understand text-based content in over 20 languages. This is still an area under research, and a focus is on reducing the models' reliance on language expertise.

Facebook's suicide prevention tools were updated in February 2018 to incorporate DeepText. Before this, scores from two N-gram based text classifiers, one focused on the text of the post, and one focused on that of the comments, were fed into a random forest learning algorithm, along with other features such as time of day, or post type. If the algorithm's output risk score was sufficiently high, the post was flagged for human review. In the update, Facebook added three DeepText classifiers as inputs to the random forest. Two re-examine the post and the comments, and one is focused on individual comments that prioritises the post for in-person intervention.[133]

DeepText has also been put to use by Instagram, initially to combat spam, and then internet trolls too. For each case, the DeepText model was trained on data specific to the issue: labelled spam content and offensive content input separately to train two models to be effective on each type of unwanted content.

4.3 AI FOR ASSISTING HUMAN MODERATORS

AI can improve the effectiveness of human moderation by prioritising content for their review

Once an automated AI system has analysed and assessed a piece of content during pre-moderation, it provides a number of 'scores' which display its level of confidence that a particular piece of content falls into a certain category. For example, an AI may score a piece of content, such as a video of a fight on a college campus, as 86% violent, 60% cruel and insensitive and 1% child abuse. If the confidence levels are sufficiently high, the content may be automatically removed or sent to a human for manual review. Using pre-moderation risk scores to triage UGC for review would improve the effectiveness of the human moderation team by allowing them to prioritise their workflow. Furthermore, prioritising content for manual review would ensure that the most damaging content is reviewed more urgently, to help

limit the exposure of harmful content to internet users before it is reviewed and taken down.

AI can reduce the challenges of moderating content in different languages by providing high quality translations

Translating text or audio information will help a moderator to assess content that is in a foreign language. This will help to increase the quality of moderation, particularly if the translation allows nuances in the language to be understood. This can be achieved using NLP and translation techniques as discussed in section 4.1.1.

Furthermore, understanding the context surrounding a piece of content will, in most cases, allow a moderator to assess the harmfulness of a piece of content more quickly, reducing the amount of time that a moderator needs to spend viewing the content. In some cases, translating words within a piece of content into the moderator's language will provide greater contextual information, reducing the amount of time that a moderator is exposed to harmful content.

AI can reduce the impact on human moderators by varying their exposure to harmful content and by controlling exposure to the most harmful elements of the content

AI could use the scoring system described above to allocate content to moderators based on what they have recently been exposed to. For example, a moderator that has previously been allocated a piece of content with a number of high scores could then be allocated content with generally lower scores. The uncertainty of the AI, reflected by the lower scores, would hopefully indicate that the content is less harmful but might still require moderation.

Object detection and scene understanding techniques can be used to protect content moderators during manual review by obfuscating the most damaging, harmful areas of flagged content. This would allow the moderators to assess the content without being exposed to its most harmful elements. If further information is required, the harmful areas be gradually revealed until sufficient evidence is visible to determine if the content should be removed or not.

Visual Question Answering techniques allow humans to determine information about the content without being exposed to it

Visual Question Answering (VQA) is a technique which allows humans to ask questions of a piece of content before they are exposed to it. This technique would benefit the moderator as it could greatly reduce the amount of content that the moderator is exposed to. For example, upon seeing that a piece of content has been flagged as potential child abuse, a moderator could ask questions such as 'Is the child clothed?' and 'Are there multiple people in the image?'. This may allow the moderator to decide that the content is harmful without ever viewing the content.

A range of previously mentioned AI techniques are required to provide VQA functionality. For example, NLP is required so that the AI can understand the question, then object detection, semantic segmentation and action recognition to allow the AI to answer the question accurately.

4.4 COMMERCIAL CHALLENGES TO USING AI TECHNIQUES

4.4.1 THE BUSINESS ENVIRONMENT DRIVES THE PRIORITISATION OF GROWING AN ACTIVE USER BASE OVER THE MODERATION OF ONLINE CONTENT

Organisations operate in a fast-paced, highly-competitive ecosystem which encourages the launch of minimum viable products and a focus on rapid growth in attracting users' attention

The internet economy has grown rapidly over the past two decades: five of the top 10 largest companies in the world (by market capitalisation) are internet-based companies,[134] as opposed to just one at the turn of the millennium. Facebook first launched in 2005 and now claims over 2 billion monthly active users, some 30% of the world's population. This very fast-paced industry is enabled by rapidly emerging innovative technologies, new business models and discerning users whose online attention is fought over intensely. Companies in this industry must focus on succeeding rapidly or they will fail.

New organisations operating in this field therefore must focus on launching products or services which are well-designed but also focussed on the features which are important without being distracted by less valuable aspects. These minimum viable products (MVPs) aim to build a base of users rapidly in order to generate the network effects necessary to sustain growth. Developing and implementing an AI-enabled content moderation system requires an investment in terms of effort and financial cost which could have been used to address other priorities of the business. Developers with the required skillsets in AI are much sought after within the technology industry and are therefore difficult and expensive to recruit.

Monetisation and profitability are generally believed to come only after a critical mass of users and content have been achieved, leading to the promotion of content which is 'clickable'

The path to success of the large internet organisations has been to generate a large user base before exploring business models which provide revenue or profits. Many start-up companies hope to emulate this approach, with a resulting focus on capturing and retaining the attention of users. Platforms which do this successfully can then expect either to be acquired by one of the dominant organisations or, in a few cases, to grow so rapidly that they become one of the internet giants themselves. In this quest for rapid growth in users and content, organisations can neglect to invest properly in the processes and safeguards which might be expected for users in an otherwise uncontrolled online experience.

The business model adopted by most of the larger, more established internet companies is to generate revenue from advertisers who pay to have their adverts placed alongside the online content viewers access. This approach similarly encourages a focus on the growth of user numbers and ways to attract and retain their attention.

These motivations for small and large organisations alike mean that, bar reputational damage, highly clickable, outrageous, addictive content, especially that with the potential to go viral, is particularly attractive for business models. While organisations also appreciate their social-responsibility obligations, they are at the same time incentivised not to over-moderate potentially harmful content if it may be valuable to their business objectives.

Internet platforms have evolved as content-sharing platforms rather than publishers with editorial responsibility

The internet has evolved from a set of mostly-static pages sharing information posted by businesses, to a richly interactive environment where individual users frequently post user-generated content about topics which may interest them and their followers. This has been accelerated by the wide accessibility of smartphones which can be used to create and edit text, images, videos and other media formats and can be shared almost instantly with the world over fast data connections.

The internet platforms behind this, in particular the social networks, have been created as content-sharing platforms to enable users to publish their own content, rather than as traditional publishers who have editorial control of their content. Therefore, the requirements for effective, scalable means of moderating online content are often considered lower priority when attempting to become a successful and established platform.

4.4.2 THE 'TECHLASH' IS GAINING MOMENTUM AND PUTTING INCREASING PRESSURE ON INTERNET ORGANISATIONS

There is a backlash against many of the dominant internet companies which is adding to the pressure to improve content moderation

In recent years there has been a growing backlash against technology companies (known as the 'techlash') which has been driven by a range of factors. Two of the main factors are the growing awareness of the public of the commercial exploitation of their data – which they have often given freely and sometimes unknowingly – and of the negative impact of harmful online content. There have been some strong headlines in the press which have brought significant attention on the ethics of social media businesses.[135] This highlights the increasing importance of online content moderation and the adverse effects which are felt by individuals and society as a whole if it is not performed adequately.

To address the range of concerns raised by the techlash, business leaders have been requested to appear before government hearings in several countries[136] and in some cases have declined to do so.[137] This has further added to the pressure by giving the impression that the industry is not adequately addressing its obligations to mitigate its negative impact on society.

The concerns of society and impact on financial performance are increasing pressure on internet companies to examine the ethics around how they operate and to put in place stronger protections for their users. This is likely to act to increase the incentive for them to implement much more rigorous processes for identifying and removing harmful online content.

4.4.3 ORGANISATIONS MUST BALANCE THE COSTS OF CONTENT MODERATION AGAINST THE BENEFITS

Implementing and operating an online content moderation process is a financial cost to commercial organisations. The appropriate process depends on the balance of the cost of getting content moderation wrong with the cost of investment to get it right as illustrated in Figure 32 opposite.

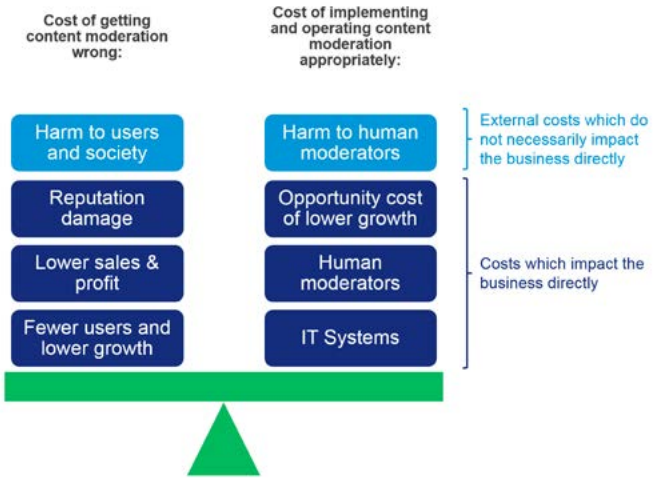


Figure 32 – Organisations must balance competing costs to determine their approach to online content moderation (**SOURCE:** Cambridge Consultants)

Content moderation addresses both legal requirements and compliance with an organisation's own policy

Content moderation can be considered in two parts:

1. **Satisfying minimum legal requirements**, for example removing child abuse material or extremist material which is defined in law. In order to avoid prosecution, reputable organisations will implement at least the minimum needed to meet their legal requirements. Organisations which operate beyond the law, for example on the dark web, are unlikely to be motivated by either legal responsibilities or any further regulation and are not considered within this report.

2. **Satisfying an organisation's own policies**, such as removing content which is not illegal but may still be harmful to users. The investment required to meet these requirements will be determined by the balance between the cost of getting the moderation wrong and the cost of implementation.

The remainder of this section considers the balance for satisfying an organisation's own policies and whether regulatory intervention may be necessary to encourage an appropriate balance for society to be reached.

Platforms have benefited from the broad immunity from liability provided by regulation

The Communications Decency Act 1996 (Section 230) in the United States and the Electronic Commerce Directive 2000 in the European Union provide immunity from liability for online service providers who publish and transmit information provided by others. Broadly speaking, these provide online platforms immunity from liability for UGC posted and shared on their sites. CDA 230 is commonly seen as "the law that matters most for speech on the web"[138] as before the introduction of this Act, platforms were liable for the content they hosted if they attempted to moderate but were not liable if they did not attempt to do so. The introduction of this legislation therefore enabled platforms to freely moderate content on their site without taking on the liability for the content, allowing online platforms and services to support rapid growth of UGC.

The cost of under-moderating content is primarily the risk to an organisation's reputation

Inadequate moderation of harmful content by organisations will result in users being exposed to that content. This has a negative impact on those users and in some cases others, such as the subjects of the harmful content. Organisations are likely to be impacted negatively by their inadequate moderation through users no longer using the service or recommending it to others, resulting in slower growth, lower revenue, lower profits and ultimately a lower corporate valuation. If this is persistent or particularly harmful then the reputation of the platform will be impacted, further driving away users and reducing subscriptions or advertising revenues. Thus, platforms moderate content as they are economically motivated to create a hospitable environment for their users in order to incentivise engagement.

In the context of the growing 'techlash', there have been growing public calls for increased regulation of online services with respect to harmful content. Therefore, in addition to driving away users resulting in impacted revenues, the larger internet companies are coming under increased pressure to improve their content moderation processes. If they are unable or unwilling to do so then they will also need to balance the impact which increased regulation may have on them. This threat may in some cases encourage improved moderation.

Conversely, over-moderating content can also have a negative impact on reputation

If an organisation over-moderates content and removes content which is not necessarily harmful then this can also be damaging to their reputation. Examples of this include the removal of the iconic 'napalm girl' photo (section 3.2.2) or removing valid user opinions which provide a balanced view on harmful content. This will also have the possible knock-on effect of reducing advertising revenue as frustrated users and advertisers migrate to new platforms.

There is, therefore, a balance which must be found in the content moderation process which ensures that the level of moderation is appropriate. The impact of over- and under-moderation is discussed in terms of false positives (material deemed to be harmful when it in fact is not) and false negatives (material which is deemed not to be harmful when in fact it is) is discussed in section 3.2.1.

Implementing and operating a content moderation process has direct costs and opportunity costs

There are direct costs associated with implementing a content moderation process and on-going costs of operating and evolving the process. The main costs for these are the time and effort in setting up the process and the cost of IT systems and staff costs for managing the process and manually moderating content as needed.

There is also an opportunity cost of the moderation process if it leads to slower development of user growth on the platform. The time and effort of managing the process and the increased complexity it adds to the overall system may have an impact on the success of the platform in a competitive business environment where achieving critical mass before competitors is key. In some cases, depending on the nature of the platform, moderating content unnecessarily could have a negative impact on the business if that content would have driven more users to access the platform. This may be the case if the material would have gone viral, perhaps because it was of a contentious nature but was not sufficiently harmful to damage the reputation of the organisation. Some platforms in a growth phase may actually benefit from having an overtly lower level of moderation, which attracts users who wish to discuss opinions more openly or share material which may fall under the moderation policies of other sites.

There are major benefits of moderating content appropriately

Although we have examined the balance of the cost of getting content moderation wrong with the cost of investment to get it right, there are significant benefits in moderating content appropriately. Users will have a more positive experience in using the platform and increasingly trust it. The organisation will build a positive reputation which will further attract more users and improve advertisers' perception of the platform.

As users gain confidence that content is moderated effectively, they are less likely to post harmful content themselves. Users with intent to post damaging material will quickly move on to other platforms. Overall this becomes a self-reinforcing reputation for a positive online community and less content moderation may be required by the organisation.

There are some external costs which may not be fully accounted for in the organisation's decision

Commercial organisations will balance the relevant costs which impact them in making a decision to invest in content moderation. However, there are external costs to individual users, to society as a whole and to human moderators who may be exposed to harmful content in the course of their duties. These external costs may not be fully accounted for by an organisation, although in many cases the risk to their reputation and their desire to be seen as a fully responsible corporate citizen will motivate them to consider these aspects.

If these external costs are not being appropriately accounted for by organisations, then there may be an argument for regulatory intervention to ensure that the external costs are appropriately addressed.

4.4.4 AI-ENABLED ONLINE CONTENT MODERATION TOOLS MUST BE ACCESSIBLE TO ORGANISATIONS OF ALL SIZES FOR THE BENEFIT OF SOCIETY

There are barriers to developing AI-enabled content moderation tools which may be greater for smaller organisations

There are three key elements to developing an AI-enabled content moderation tool. Each of these can potentially act as a barrier to an organisation being able to develop such a tool:

1. Algorithms and the skills to implement them
2. Data sets to train AI systems
3. Computational power for the training stage of development

These are described in turn in the following paragraphs, but overall larger players benefit from their ability of finance to invest in development, access to data and economies of scale.

1. **Algorithms are generally available, but there is a shortage of staff skilled in implementing them which could disproportionately impact smaller organisations**
   The open and collaborative nature of AI research means that many algorithms and approaches have been published and tools and techniques are available as open source. However, skilled staff are required to understand the design choices and optimisations required to apply these. There is a widespread shortage of suitably qualified staff with skills in AI development due to the level of investment in AI developments in the commercial sector. Recruiting and retaining skilled staff may be easier for

large, established technology businesses who can attract staff and cross-subsidise their costs with profits from other lines of business. Despite this, the rise of third-party content moderation services and commercial off-the-shelf moderation solutions may address the skill deficit by providing smaller organisations with the appropriate technologies and access to the skilled staff required to implement them.

2. **Data is required for training AI systems, which could act as a barrier to some organisations, but AI techniques can also be used to generate more data**
   The availability of sufficient high-quality data needed to train an AI-enabled content moderation system may act as a blocker to organisations which do not already possess a suitable set of data. However, there are tools that may help to reduce the amount of training data required, although not to remove the need altogether.

   AI techniques, such as GANs, can be used to generate additional data successfully for this purpose, as discussed in section 4.2. These take a data set and augment it by creating additional examples using the characteristics of the original dataset. Although this may appear to be counter-intuitive, as AI is creating the data that is then used to train it, it has been shown to be an effective approach in practice.

POLICY IMPLICATION

The sharing of datasets applicable to identifying harmful content between platforms and moderation service providers will help to maintain standards and should be encouraged. Data Trusts (as identified in the UK Government's recent AI review[139]) may offer a suitable framework for this. Relevant data held by public bodies, such as the BBC, could be contributed for the benefit of society by enabling a comprehensive dataset to be made available and kept up to date with evolving categories of harmful content and formats.

3. **Access to computational power is unlikely to act as a barrier**
   Computational power is required during the training stages of developing an AI system and during the inference stage when the AI is operating. As cloud computing now provides an established means of leasing computational power as and when it is needed, this is unlikely to be a significant barrier to organisations. Large organisations which provide cloud computing services may be able to subsidise their own access to this resource to reduce this barrier for themselves, but the competitive nature of this market and the availability of low-cost off-peak computing resource will limit this competitive advantage. Hardware designed specifically for AI is also increasingly available, in the cloud and on-device, this will continue to reduce in price and become more accessible for a range of users.

Offering moderation as a service appears to be a viable business model which facilitates access to all players

Some organisations also offer content moderation as a service to platform providers. This means that it is not necessary for a platform provider to develop an in-house capability for moderating content. Organisations which offer this service can benefit from economies of scale when developing technical expertise in recognising harmful content, whilst online platforms benefit as they can contract out this service.

Content moderation services are available for pre-moderation by AI and post-moderation by humans:

1. **API access to AI-enabled content pre-moderation**
   Access to systems which can identify harmful content is available through APIs. An API is an interface which allows easy porting of functionality between applications. Typically, the system is provided with a URL to the content to be checked and the system returns its interpretation of the content and whether it should be flagged as harmful. These systems often provide high levels of functionality and interpretation of the content, such as classifying the subject of the content, identifying the number and location of people and other objects within an image, and analysing the sentiment or emotion of the content. However, considering the context of UGC can be more difficult as the amount and type of metadata available depends on the features of each specific platform (see section 4.1.2) These services are typically suited to the pre-moderation stage in the moderation workflow (see section 3.1.3) but do not provide human moderation services.

   These content analysis and moderation tools are provided by a range of companies, including many of the large internet companies such as Amazon[140], Google[141] and Microsoft[142], and many smaller companies such as Xmoderator[143] and clarifai[144].

2. **AI-enabled and human content moderation as a service**
   There are numerous organisations which offer a fully managed service for content moderation. These organisations typically manage the full content moderation workflow for a platform, including pre-moderation using AI techniques and teams of humans to review content marked for further attention. These services provide a premium

service over pre-moderation alone and typically offer a high degree of control and customisation to the types of content which are permitted.

Examples of these organisations are Crisp Thinking[145], Two Hat Security[146] and Besedo[147].

The availability of moderation as a service from both large internet companies and smaller, specialised organisations indicates that there is a functioning market for providing content moderation services. Competition in this market will drive further innovation, advances in techniques, a reduction of prices and increase in performance which makes the service commercially viable to range of different size players.

POLICY IMPLICATION

The availability of online content moderation services from third-party providers should be encouraged. This will help to ensure that services are accessible to platforms of all sizes and will encourage the use of AI and automation techniques to increase the performance and effectiveness of content moderation.

Market intervention may be needed to ensure that high-performance content moderation is available to the ecosystem for the greater good of society

Some anecdotal research by others indicates that the performance of commercially-available content moderation services varies between services and for different types of content. For example, research published by Towards Data Science[148] tested systems with 120 images across five categories of harmful content and compared the accuracy of categorising these as "safe for work" and "not safe for work". The research showed that systems were uniformly good at categorising images showing explicit nudity, but performance was variable for detecting gore as illustrated in Figure 33 below.

The reason for this variation is likely to be the amount of training data in these categories that was used to develop the systems and the level of testing and optimisation which had been carried out on these categories. Variations in performance across categories may not be clear to organisations procuring services for content moderation and any published accuracy statistics may not be comparable between systems. Statistics may also only apply to certain categories which may not be the categories of greatest importance to the purchasing organisation.

The international nature of the commercially-available content moderation services is also likely to lead to a lack of sensitivity to national or cultural differences. The importance of this is discussed in section 3.2.3.

POLICY IMPLICATION

To ensure appropriate levels of internet user protection, it is important to be able to understand the performance of AI-based content moderation by individual platforms and moderation services across categories. This is to confirm that these deliver appropriate moderation and that they are evolving with the expectations of society and relevant national or cultural sensitivities.

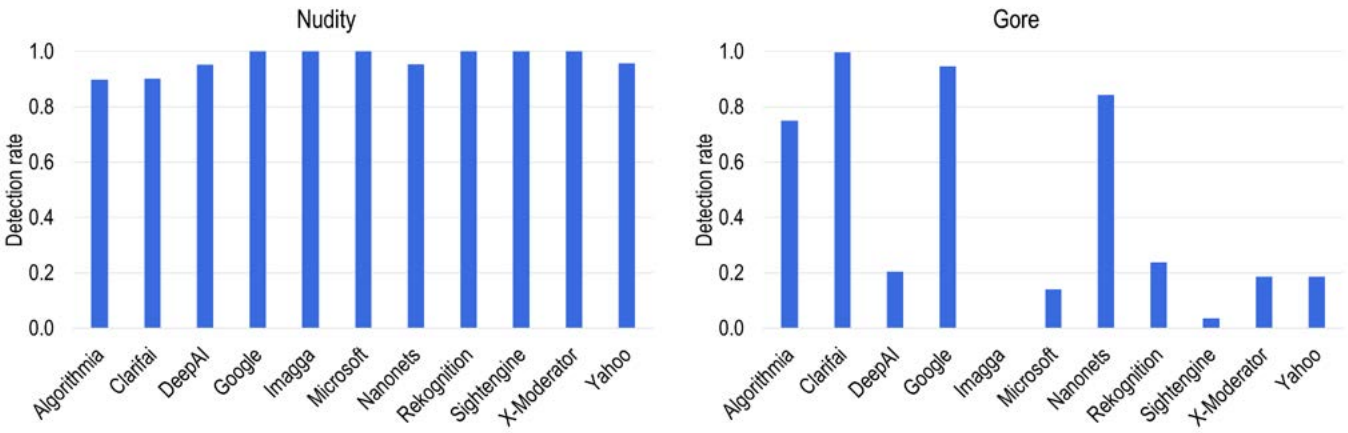


Figure 33 – The performance of content moderation services varies for different types of harmful content (**SOURCE:** Cambridge Consultants' representation of data from towardsdatascience.com)

5 POTENTIAL IMPACT OF AI ON PROMOTING SOCIALLY-POSITIVE ONLINE ENGAGEMENT

Human interaction has developed over thousands of years. Throughout this time, humans have learnt to interpret cues such as body language and facial expression to enable more advanced, socially intelligent interactions. When online, however, these cues are removed as text and voice are the predominant methods of communication and even video calling often fails to convey cues as effectively as a face-to-face interaction.

As a result, online communication which often lacks emotional expression can result in harmful or negative interactions, particularly between strangers. The 'online disinhibition effect' offers an explanation as to why humans may behave worse online.

Encouraging socially positive engagement is a benefit to platforms and users alike. It is an effective technique for platforms to reduce the amount of content moderation that they have to implement, as less harmful content is posted, and users can benefit from a more hospitable environment online.

Platforms are trying to address this with methods of promoting socially positive engagement online. In practice, most of these are quite coarse filters, such as email verification to reduce access for online trolls. However, these techniques have been shown to reduce the amount of objectionable content that appears online.

Promoting socially positive engagement can be an effective tool for reducing harmful online content, but it is important to consider the spectrum of internet users and their motivations for posting harmful content. Whilst techniques such as nudging may encourage some internet users to sense-check before posting harmful content online, promoting socially positive engagement is unlikely to inhibit truly malicious actors such as those who wish to incite hate through terrorist propaganda. As such, the effectiveness of nudging techniques will vary depending on the user's intentions and the category of harmful content.

We have identified some AI techniques that can be used to further eliminate the amount of harmful content that appears online and other AI techniques that can allow users to moderate the content that they are exposed to themselves.

5.1 PEOPLE BEHAVE DIFFERENTLY ONLINE TO IN-PERSON

The online disinhibition effect explains why internet users may act maliciously online

The proliferation of harmful, objectionable and undesirable content online can be attributed to a number of factors. The online disinhibition effect explains why many people communicate negatively much more frequently than they would in person.[149] The online disinhibition effect is caused by a combination of factors such as:

- anonymity
- asynchronous communication
- an empathy deficit

Anonymity associated with online interactions can result in internet users feeling empowered to express themselves more freely. Whilst this can result in internet users being more open and willing to share personal feelings in positive ways, it can also encourage internet users to act more maliciously online.

Asynchronous communication, meaning that recipients do not necessarily receive the message and respond immediately, can be beneficial as it allows internet users to formulate their thoughts before replying. However, it can result in harmful online behaviour as internet users do not see the immediate consequence of their actions in the same way as they do in face-to-face interactions.

Online interactions lack empathy as internet users have no visibility into how others are feeling. The empathy deficit can cause internet users to act with a lack of constraint and compassion online as they do not see the effect their harmful content has on other individuals.

These factors cause significant complications to human interactions when performed online. Encouraging socially positive engagement and interactions online will likely require the development of more advanced online 'cues' such as emojis which aim to integrate emotion into online interactions. Introducing emotion into online interactions will likely result in fewer harmful interactions by reducing the empathy deficit in which victims of bad behaviour are reduced to text on a screen.

5.2 THE BENEFITS OF ENCOURAGING SOCIALLY POSITIVE ENGAGEMENT

Encouraging socially positive engagement online reduces the need for content moderation

Whilst content moderation focuses on detecting and removing UGC which has already been uploaded, promoting positive engagement aims to incentivise positive behaviour whilst disincentivising bad behaviour. Content moderation and the promotion of positive engagement are interlinked: content moderation focuses on removing harmful content, while promoting positive engagement discourages the user from posting harmful content in the first place. Promoting positive engagement can be considered a form of content moderation as it aims to reduce the quantity of uploaded bad content at the source by encouraging good user behaviour.

Whilst content that breaches community guidelines must be moderated, online platforms which aim to provide a positive environment are often littered with objectionable content which may not breach its rules or guidelines but does, however, create a negative or unpleasant environment for its users.

Online content moderation will never be perfect so harmful content should be discouraged

Contextually complex examples like bullying can be much harder to detect than content which is clearly harmful even without reference to its context. For example, detecting bullying online requires an understanding of the relationship between two users, their age, the number of exchanged messages, the nature of their connection as well as previous interaction history and shared connections. An effective moderation process should differentiate between sarcastic remarks between friends and true cases of online bullying. These nuances of human language and online interaction introduce grey areas to the content moderation process.

Promoting socially positive online engagement can be an important tool for reducing harmful content that is inherently complex to moderate. Encouraging positive online engagement and interactions can be an effective approach to reducing the content moderation requirement of contextually difficult material. Organisations which operate online platforms should explore techniques to reduce harmful content at the source. This would be especially beneficial for content types which may cause significant reputational and social damage having bypassed the content moderation process.

5.3 COMMON NON-AI APPROACHES TO PROMOTING SOCIALLY POSITIVE ONLINE ENGAGEMENT

Many online platforms are investigating ways in which users can be encouraged to engage positively with the platform and its users to help minimise the quantity of harmful content uploaded to their platforms. Promoting positive engagement can be performed in many ways, using a number of incentives and techniques. Techniques range from practical measures such as using multi-factor authentication to verify a user's identity to automated, rule-based algorithms.

Multi factor authentication reduces online anonymity

A common approach to reducing the amount of uploaded spam is to require multiple factor authentication (MFA) in which each account must be verified using, for example, a mobile number. This technique is intended to restrict each user to a single account, reducing the number of spam and duplicate accounts and thus the quantity of spam content. Furthermore, MFA reduces the ability of opening fake accounts which reduces online anonymity and thus reduces the online disinhibition effect. Forcing internet users to link their accounts to real identities can be an effective tool for not only discouraging spam but also for encouraging positive interactions online.

However, there can be difficulties when inactive contact details are re-used by others, for example when a disused mobile phone number is recycled by operators and given to a new subscriber or when an inactive email address is made available again by the email service provider. This creates a limitation in this attempt to prove identity and can itself lead to difficulties where personally-sensitive emails intended for the previous user are received by the new user of an email address.

Demonetisation can discourage posting of harmful content

After a highly influential account posted a video of an apparent suicide victim in Japan, YouTube have announced that human moderators will review content of big partner accounts, turning off monetisation on any specific video they deem "not advertiser friendly".[150] YouTube hopes demonetising negative content will disincentivise users from posting such content while maintaining appropriate levels of freedom of speech. Whilst this technique may encourage users to sense check their content prior to uploading, it is unlikely that such a tool will disincentivise truly malicious actors. Furthermore, this technique is only applicable to previously monetised accounts and as such cannot be used to disincentivise smaller, potentially more damaging user accounts.

Censure can be effective in reducing harmful interactions

Another common form of promoting socially positive engagement is to employ a form of social punishment for bad behaviour. League of Legends, an online game which boasts over 80 million active players, implemented a new strategy to promote positive interactions. Bad users who repeatedly interact negatively on the platform are flagged and their accounts are restricted to a limited budget of interactions. Those users must then decide if they want to utilise their limited interactions in a positive manner to engage with others and improve the platform or not. Players who modify their behaviour to be more positive can be released from the limited chat mode. The publisher of the game, Riot Games, highlighted that bad language as a whole dropped 7% and that positive interactions increased.[151]

Algorithmic curation tools can reduce harmful interactions online

Online platforms curate content on their sites using algorithms. These algorithms suggest content which they believe the user will engage with in an attempt to drive interaction through comments, views, likes or shares. These algorithms prioritise content for users based on the interaction of other users with that content. For example, Facebook prioritises content that will "spark conversations and meaningful interactions",[152] Twitter assesses behavioural signals to determine and prioritise positive interactions[153] and YouTube recommends popular or trending videos. Whilst these algorithms aim to increase user interaction for monetary purposes, they can also be an effective tool for promoting a positive online experience by giving greater exposure to positive content.

These curation algorithms can in themselves promote content which is harmful or can direct a conversation towards a more harmful direction by focusing on the content which gains more interaction. More positive content does not necessarily drive as much interaction and so may not be promoted as effectively as a result. More advanced versions of these algorithms, such as those which allow users themselves to up-vote or down-vote content can in similar ways leads to a bias towards more negative content being promoted online.

The AI techniques described in section 4.1 can be used to assess the sentiment and level of harmfulness of content. This can be used as an input parameter to the algorithms which decide which content to prioritise. As the AI techniques improve in their capability, it will become increasingly beneficial to use these metrics within the prioritisation algorithm.

5.4 IMPACT OF AI ON TECHNIQUES TO PROMOTE SOCIALLY POSITIVE ONLINE ENGAGEMENT

AI techniques can help inform users about content before they see it

Automatic Keyword Extraction (AKE) analyses for key words, key phrases and key segments that can describe the meaning or sentiment of a document or piece of content. AKE is an important aspect of information retrieval which underpins web search engines, allowing relevant content to be displayed by linking database content to a search query.

AKE can be used to help identify content that may be harmful in general or specifically harmful to vulnerable groups online. This could allow users to understand the content of a post, prior to viewing, and then to make an informed choice about whether they wish to view the content. For example, this technique can be used to restrict access to content that promotes Non-Suicidal Self-Injury (NSSI). Exposure to this content in online forums and on social media has been found to trigger NSSI behaviour in individuals.[154] Hashtags used on NSSI content are often ambiguous and difficult to identify. Currently, searching for known harmful tags on Instagram results in a pop-up prompting users to get support, or see posts anyway. However, the app relies on users to report these hashtags and therefore does not consider alterative spellings. Using AI, in the form of AKE, for automation would simplify the process and allow it to be done at scale.

AI nudging techniques can discourage users from posting harmful content by prompting them to think again

AI can be used to encourage more socially positive content. For example, when a user posts new content, sentiment analysis can be used to judge if the text contains content that is offensive or harmful. If so, several steps could be taken to discourage the user from posting.

A simple prompt may be raised asking the user if they really want to post the content. Alternatively, action delay could be used to impose a set waiting period before posting to help prevent impulsive posts. A YouGov survey[155] in the US found that 28% of users have posted something on social media that they regret. 65% of these people said the reason they posted was either than they didn't properly consider a response or that they responded in the heat of the moment. Nudging uses subtle techniques to create incentives for some choices and not others by manipulating the choice architecture to incentivise the most beneficial decision. A small nudge through an alert or delayed posting may have allowed these users to re-consider, thereby reducing the number of negative posts.

Through investigating the comments sections below online articles, researchers at Cornell University's Department of Information Science state that there are two main triggers which determine when a user may begin posting harmful content: the context of the exchange and the mood of the user. By collecting comment section data, the researchers have developed an algorithm that predicts with 80% accuracy when a user is about to become abusive online.[156] Researchers suggest that introducing a delay to the speed in which users can respond can change the context of the exchange. By introducing a delay, users are more likely to think about their comment and the effect it may have on the victim, increasing empathy and reducing the likelihood of a 'bad' interaction.

AI techniques can nudge users to post less negative content by suggesting alternatives

AI techniques can be used to suggest less harmful ways for users to post their message. Unsupervised text-style transfer could be used translate offensive language within a post into non-offensive forms and provide the user with the option to use the clean version. Research[157] published by IBM in 2018 proposed using an RNN encoder to parse an offensive sentence and compress the most relevant information. This is then read by an RNN decoder to generate a new sentence. The resultant sentence is then evaluated by a CNN classifier.[158] This determines whether the output was correctly transferred from the offensive to non-offensive style. To further assess the success of the style transfer, the output is back-translated from non-offensive to offensive. This resultant offensive sentence is compared to the original to check that the meaning has not been lost.

This enables a suggested rewording to be created which preserves the sentiment in a more sophisticated way than simply removing expletives or harmful words. However, the researchers acknowledge that in some examples the technique did not suggest a meaningful alternative and so further work is required before this approach can be used in a real system.

TEXT SOURCE	ORIGINAL TEXT	TEXT SUGGESTED BY AI
Reddit	"for f**k sake, first world problems are the worst"	"for hell sake, first world problems are the worst"
Reddit	"what a f**king circus this is"	"what a big circus this is"
Twitter	"I'm back bitc**s!!!"	"I'm back bruh!!!"

Figure 34 – An AI model to suggest less offensive language can help in some cases (**SOURCE:** Cambridge Consultants' representation of extract from IBM research paper[159])

AI-powered chatbots can prompt users to report harmful content

Chat bots can be used on social media to encourage other users to be more aware of negative content online and remind them of their social responsibility. In 2015, a study[160] was conducted on the role of nudging techniques within the role-playing game League of Legends. The study found that when a player was prompted to report another player for abusive behaviour, they were 16 times more likely to do so. Chat bots could be used to this effect by encouraging users to take responsibility for their own posts and reporting abusive users.

AI ALGORITHMS CAN HELP USERS TO COORDINATE

Researchers at Yale University have conducted a number of experiments investigating strategies to encourage more positive and cooperative interactions online. The experiment investigated the effect of introducing random AI bots to networks of human nodes to understand and quantify its effect on the networks' ability to collaborate to complete a simple task. The experiment demonstrated that an unpredictable and unintelligent AI bot that behaves randomly can nudge users to coordinate actions with others to accomplish simple tasks more quickly.[161] The results highlight that bots that were placed centrally and randomised their decision 10% of the time outperformed all the human networks. These collaborative human-AI networks solved the coordination game more frequently (85% vs. 67%). Furthermore, the median time required was reduced from 232 seconds to 103 seconds.[162] The researchers understand that the experiment is limited in its complexity and hope to include more complex and realistic collaborative tasks in their research to demonstrate how human networks (such as social networks) may be encouraged to collaborate using AI.

AI tools for reducing online disinhibition can improve online interactions

AI chatbots can be used to highlight negative content to users and prompt an improvement in their online behaviour. One experiment[163] used Twitter bots to reply to internet users who have harassed black users with racist abuse, an example of which is shown in Figure 35. The experimental chatbots, which appeared to others to be real users, automatically replied to racist tweets and demonstrated that users were much less likely to post racist abuse following such a chatbot intervention.

The results published in the journal Political Behaviour demonstrated that negative responses to racist tweets impacted their likelihood, by reducing the empathy deficit of those posting racist abuse. This technique may therefore be valuable for online platforms to nudge users into less harmful interactions but consideration needs to be made of how appropriate it is to use of chatbots purporting to be real users.

Nudging techniques may be useful for reducing harmful interactions, but careful consideration is required to ensure it is implemented correctly

Whilst nudging techniques can be effective tools for encouraging good decisions and positive interactions, many argue that their use by platforms to change citizens' behaviour, even if these changes may be beneficial to the individual, violate the privacy and integrity of the individual.[164] Additionally, the problem persists of a platform making a value judgement about what is beneficial and how they should balance the definition of their community standards with the freedom of expression of their users.

Furthermore, whilst nudging techniques may help encourage online interactions to be positive, excess nudging could be detrimental as it could stifle constructive debate by discouraging users from challenging accepted cultural norms.

The benefits of nudging are apparent. However, it is important that users still maintain a freedom of choice. Choosing the non-desirable option, from the platform's perspective, should be easy for the user. This is backed up by Nudge, a book written by economist Richard Thaler and Harvard Law School Professor Cass Sunstein which investigates the psychology behind nudging. They argue that manipulating the choice architecture through nudging is legitimate as long as people are still able to make their own decisions sufficiently easily.

KEY FINDING

Promoting socially positive engagement can reduce the need for content moderation

The most notable benefit of promoting socially positive online engagement is the improved online environment experienced by internet users. By disincentivising them from posting harmful content, users are less likely to create malicious content and interact negatively, thereby improving the online environment for everyone. Similarly, incentivising good behaviour is also a tool that can be used, and much research is being focused into using automated tools to categorise and prioritise good content to reduce exposure to harmful or negative content. However, further research is required to accurately define what constitutes positive content such that the success of moderation and algorithmic design policies can be measured.

Figure 35 – Chatbot tweets in response to a racist message have been shown to reduce the subsequent number of racist tweets (**SOURCE:** Kevin Munger,[165] used with permission)

6 POLICY IMPLICATIONS

The policy implications highlighted below are to inform stakeholders from government, regulators, internet sites and services, and stakeholder groups of some of the possible areas of concern around AI-based content moderation.

These implications should be considered in the broader context of the evolving landscape of content moderation generally. Online sites and services offering transparency on their definition and treatment of harmful content will help to protect users and ensure that they, and other stakeholders, will gain confidence in the standards of the platform. These sites and services should endeavour to establish the informed consent of users about the type of content which may be encountered on the platform and offer appropriate warnings. This will ensure that users are able to make informed decisions about whether the platform is appropriate and, where appropriate, understand how to flag inappropriate content.

There is also promising research into the impact of techniques used to encourage socially positive online engagement. Should such techniques prove to be widely applicable, this could reduce the reliance on online content moderation systems and reduce the amount of harmful content posted for some categories.

	CONTEXT	POLICY IMPLICATIONS
1	AI-enabled online content moderation tools are proving to be a necessary part of online platform providers' response to harmful online content and tools must be accessible to organisations of all sizes. Data is required for training AI systems, which could act as a barrier to some organisations. See section 4.4.4 for more information.	The availability of online content moderation services from third-party providers should be encouraged. This will help to ensure that services are accessible to platforms of all sizes and will encourage the use of AI and automation techniques to increase the performance and effectiveness of content moderation.
2	The data used to train AI-enabled online content moderation services will evolve as different user behaviours come to the fore. Sharing data between organisations would assist the entire sector in adopting a collaborative and coordinated approach to managing harmful content. See section 4.4.4 for more information.	The sharing of datasets applicable to identifying harmful content between platforms and moderation service providers will help to maintain standards and should be encouraged. Data Trusts (as identified in the UK Government's recent AI review[166]) may offer a suitable framework for this. Relevant data held by public bodies, such as the BBC, could be contributed for the benefit of society by enabling a comprehensive dataset to be made available and kept up to date with evolving categories of harmful content and formats.
3	Users and other stakeholders will need to feel confident in the performance of any AI-based content moderation approach, especially in the early days of technology adoption. See section 3.2.3 for more information.	It is important to build public confidence that any potential sources of bias in AI-based content moderation are understood and appropriate steps taken to mitigate them. This may be through auditing and calibrating datasets to understand how representative they are of the diversity of individuals in society; or by establishing a testing regime for AI-based content moderation systems.
4	It is not always possible to fully understand how an AI system makes its decisions or how effective it is in meeting users' expectations of protection. See section 4.4.4 for more information.	To ensure appropriate levels of internet user protection, it is important to be able to understand the performance of AI-based content moderation by individual platforms and moderation services across categories. This is to confirm that these deliver appropriate moderation and that they are evolving with the expectations of society and relevant national or cultural sensitivities.

APPENDIX A: SUMMARY OF KEY AI TECHNOLOGIES

There are many different technical terms frequently used in the field of AI. Although it is not possible to provide a comprehensive guide to all of them, this Appendix provides a brief description of the main concepts.

We have adopted the following structure to provide guidance on the relationships between these terms, but note that there is on-going discussion in the AI community on the definitions and relationships of these terms:

1. LEVELS OF AI CAPABILITY	▪ Narrow AI ▪ General AI ▪ Artificial super-intelligence
2. COMMON TASKS FOR AI	▪ Search and planning ▪ Perception ▪ Machine translation ▪ Agents and actions ▪ Generation of audio, images and videos ▪ Natural language understanding (NLU) ▪ Machine vision
3. AI BUILDING BLOCKS	▪ Natural language processing (NLP) ▪ Sentiment analysis ▪ Object detection ▪ Semantic segmentation
4. AI APPROACHES AND CONCEPTS	▪ Machine learning (ML) ▪ Supervised learning ▪ Unsupervised learning ▪ Semi-supervised learning ▪ Reinforcement learning ▪ Domain adaptation and transfer learning
5. AI TECHNOLOGIES AND ALGORITHMS	▪ Artificial neural networks (ANN) ▪ Convolutional neural networks (CNN) ▪ Recurrent neural network (RNN) ▪ Long short-term memory (LSTM) ▪ Generative adversarial network (GAN) ▪ Dynamic programming ▪ Graph networks ▪ Bayesian inference ▪ Support vector machine (SVM) ▪ Gaussian processes ▪ K-means clustering ▪ Capsule networks

A.1 LEVELS OF AI CAPABILITY

Narrow AI

A narrow AI is an autonomous system that can carry out a single action. For example, Deep Blue can only play chess; self-driving car systems cannot operate autonomous aircraft and autocorrect systems cannot write novels independently.

General AI

A general AI is a system that can apply intelligence to any problem. Some consider these to be as intelligent as humans. These might be systems more akin to those portrayed in science fiction: KITT in Knightrider, Marvin the Paranoid Android of The Hitchhiker's Guide to the Galaxy or more ominously, Skynet in the Terminator franchise.

Artificial super-intelligence

An artificial super-intelligent system would meet or exceed human intelligence, creativity and social skills. With a vastly superior memory base, intellect and ability to multitask, a super-intelligent system would be able to easily outcompete humans in any way it may choose. If this occurs then a superintelligence would hugely improve the welfare of humans around the world, but also potentially cause massive destruction to humans and the environment. The concept of superintelligence is often associated with the notion of a technological singularity, where an AI system triggers runaway technological growth by continuously improving and replicating itself.

A.2 COMMON TASKS FOR AI

Search and planning

Search and planning are concerned with reducing multiple available options, be they solutions, paths or goals, down to one or a subset. Search can be performed with or without heuristics to guide selection. In most cases an algorithm will aim to perform such that an exhaustive search is avoided.

Perception

Perception deals with enabling machines to understand and interpret the inputs used by humans, covering **machine vision**, **machine hearing** and **natural language understanding**. This relies on many of the building blocks described below, such as semantic segmentation, object class and detection and natural language processing. The majority of research in these areas falls within machine learning due to the availability of data for training and researching algorithms, and the fragility of alternative approaches in the presence of varied real-world data.

Machine translation

Translation between languages by machines has long been an objective for AI because it is a highly valuable application which is complex and relies on skilled humans to perform with a high level of accuracy and reliability.

Agents and actions

Agents in AI are concerned with acting and interacting with dynamic environments. The agent must operate over time, cycling between observing and acting. This is one of the oldest tasks in the field and is often the one the public things of when hearing the term AI. The definition of an agent and its actions depends on the boundary between the agent and its environment – a poker playing agent could be capable of making choices for each hand played while non-AI control systems deal with picking up and laying down cards and money, or it could do all playing choices and object interaction.

Generation of audio, images and videos

The use of AI to generate convincing outputs of audio, image and video content has been gaining ground in recent years due to rapid progresses made in the research and application of generative adversarial networks (described below). This is valuable to generate high-quality media cheaply and effectively and can be used to augment training data for machine learning techniques. As an application of AI, it raises significant concerns around the potential of machines generating convincing media as society grapples with the implications of 'fake news' and its ability to improperly influence individuals' opinions.

Natural language understanding (NLU)

NLU focuses on the comprehension of text, often with a specific task in mind, such as summarising an article. This can be used together with natural language generation which uses the understanding to produce an output, forming a complete system that does not require human supervision. Chatbots on websites that can respond to customers' issues are a good example of this.NLU is a major area of current research. Use of deep learning models has enabled significant progress in the field compared to using purely rule-based systems, which struggle with the complexity of human language.

Machine vision

Machine vision encompasses all technology (hardware and software) that enables information collection from images. There is an image capturing step by a camera or sensor, then an image processing to extract the necessary data. A common use of machine vision is in barcode reading, and machine vision is also a key area in robotics and autonomous vehicles research.

A.3 AI BUILDING BLOCKS

Natural language processing (NLP)

NLP is concerned with machine reading and understanding of natural language. This covers text processing, retrieval, summarisation, translation and question answering, among many others. The majority of research in this area uses machine learning techniques.

Sentiment analysis

When analysing natural language, in either text, audio or video form, sentiment analysis refers to the understanding of the intention, emotions or even body language behind the words used. This enables a much clearer perception of meaning than can often be extracted from the words alone.

Object detection

Object detection refers to the identification of specific types of objects within an image. Learning the properties of these object classes enables an AI system to detect the presence and location of an object belonging to that class.

Semantic segmentation

Semantic segmentation is the process of analysing an image and identifying which pixels within it belong to which object classes identified in the image.

A.4 AI APPROACHES AND CONCEPTS

Machine learning (ML)

ML is a general framework and broad family of techniques that enable computer systems to improve their performance on a task with experience. The experience is typically built from examples of task data, but can also be simulated or real-world experience. ML is by far the largest field within AI and is the focus of the most active research. Increasingly other subfields of AI are being tackled with ML methods. Prominent applications of ML are object recognition, speech recognition, clustering and recommender systems.

Supervised learning

A learning strategy in which each example used for training consists of input data and label pairs. The system has access to the correct label and must learn the mapping from input to label.

Unsupervised learning

In contrast to supervised learning, the examples used for training consists of input data with no label. The system must therefore learn the underlying structure of the data without labels to guide it.

In unsupervised learning, the structure or probability density learned may itself be the desired output of the process, for example in clustering the structure enables data points to be assigned to subsets. Alternatively, it can be a part of further processing or learning, for example in feature learning, the learned structure of the data becomes the input to a supervised learning process.

Semi-supervised learning

Often the situation arises where a large set of unlabelled data exists, with a smaller subgroup of labelled data. Semi supervised learning methods utilise both labelled and unlabelled data. Using semi-supervised learning it is possible to combine the advantages of working with a small labelled dataset to guide the learning process and a larger unlabelled dataset to increase the generalisability of the solution. It is an important approach due to the expense of labelling certain types of data, such as in the medical imaging field

Reinforcement learning

In reinforcement learning, feedback is gathered by the system through interaction with its environment. Some interactions generate positive or negative rewards. The system must select actions to maximise the rewards.

An example of this might be in learning to play a game such as Pac-Man. A reinforcement learning system would experiment with its play with the goal of maximising its score and would therefore learn what actions led to positive and negative rewards.

Domain adaptation and transfer learning

Domain adaptation refers to the ability of a machine learning system to be applied to a different but related type of data from its training set. Transfer learning refers to the process of a system learning to adapt itself to the different domain. One such example is a system which has been trained to identify a specific object in an image can learn to identify a different object.

A.5 AI TECHNOLOGIES AND ALGORITHMS

Artificial neural networks (ANN)

ANNs comprise of many layers of interconnected nodes, in a similar way to neurons in animal brains. These networks learn by being considering training data which has the effect of reinforcing or weakening the links between nodes. Applied over very large numbers of neurons, these networks learn to identify specific features or characteristics of their input data.

Convolutional neural networks (CNN)

A CNN is an ANN in which each layer acts as a filter for the following layer. Each successive layer of a CNN recognises more complexity and abstract detail of the input data than the last.

Recurrent neural network (RNN)

An RNN processes an ordered series of data to identify complex patterns across multiple steps. This is the approach that is generally used for natural language processing as it recognises the relationships between arrangements of words and sentences.

Long short-term memory (LSTM)

A type of RNN which is well suited to classify, process and predict time series with time lags of unknown size. It is good at learning from sequences and predicting rare events.

Generative adversarial network (GAN)

Two neural networks are pitted against each other in order to generate outputs which are convincingly similar to an input set of data. One AI system creates examples which are provided to a discriminator system, along with the occasional real example, to decide whether this is a convincing example. This feedback loop allows both systems to improve, become a better creator of new data and a better discriminator between created data and the input data.

Dynamic programming

Dynamic programming allows a complex problem to be solved by incrementally breaking it down into a collection of simpler problems. Once the simpler problems have been solved then their solutions are stored so that they can be retrieved quickly in the future. The solution to the overall problem can be found more quickly by looking up the solutions to the constituent problems and combining them.

Graph networks

Graph networks are a type of architecture of neural networks which encodes the relationships in neural networks. It is an area in which a lot of research is currently focussed, and significant progress can be expected.

Bayesian inference Bayesian inference allows the prior understanding of humans to be encoded into the machine learning model, independently of what the training data may indicate. This allows better machine learning outputs to be produced when training data is limited and there is already some understanding of the problem.

Support vector machine (SVM)

Support vector machines are a set of supervised machine learning methods that can be used for classification, regression and outlier detection. For classification tasks, SVMs learn optimal hyperplanes (multidimensional lines) that separate training examples based on their dimensional representation. New examples can then be categorised by considering their dimensional representation with respect to the separating hyperplane.

Gaussian processes

In machine learning, a Gaussian process is a statistical approach to measure the similarity between an observed data point and training data. It is based on Gaussian distributions of data across many different variables and so provides uncertainty information based on the distribution of data.

K-means clustering

K-means clustering is a statistical technique to group data points into a smaller number of clusters, such that the difference between each data point and the mean of its cluster is minimised. This is a computationally expensive problem which is valuable in grouping observations. It is related to the 'k-nearest neighbour classifier' machine learning technique which identifies the closest training examples for an observed data point.

Capsule networks

Capsule networks are an architecture design that have shown considerable promise. A capsule is a group of neurons that are trained to recognize the presence of visual entities and encode their properties into vector outputs. This design feature preserves detailed information about an object's or object-part's location and its pose within an image. Within the framework of neural networks, several capsules can be grouped together to form a capsule layer. This approach allows the relationship between different entities to be learnt, for example the relative positions of facial features.

APPENDIX B: ABBREVIATIONS

The following abbreviations for technical terms are used in this report:

AGI	Artificial General Intelligence
AKE	Automatic Keyword Extraction
AI	Artificial Intelligence
ANI	Artificial Narrow Intelligence
ANN	Artificial Neural Network
ASI	Artificial Super Intelligence
BoW	Bag of Words
CNN	Convolutional Neural Network
CPU	Central Processing Unit
GAN	Generative Adversarial Network
GIF	Graphics Interchange Format
GIFCT	Global Internet Forum to Counter Terrorism
GPU	Graphical Processing Unit
HOG	Histogram of Oriented Gradients
LSTM	Long Short-Term Memory
MEC	Multi-Access Edge Computing
MFA	Multi Factor Authentication
MIMO	Multiple-Input and Multiple-Output
ML	Machine Learning
MVPs	Minimum Viable Products
NAS	Neural Architecture Search
NCMEC	National Centre for Missing and Exploited Children
NLP	Natural Language Processing
NLU	Natural Language Understanding
RL	Reinforcement Learning
RNN	Recurrent Neural Network
SON	Self-Organising Network
SVM	Support Vector Machine
TPU	Tensor Processing Unit
UGC	User Generated Content
URL	Uniform Resource Locator
VQA	Visual Question Answering

ENDNOTES

1 For example, the number of Stories (a feature in which uploaded photos and videos automatically vanish after 24 hours) posted on Instagram has risen from 100 million per day in late 2016 to 400 million per day in September 2018. Data extracted from https://www.socialreport.com/insights/article/115005343286-Instagram-Stories-Vs-Snapchat-Stories-2017-Statistics and https://www.omnicoreagency.com/instagram-statistics/

2 The ability of AI to comprehend human language

3 The ability of AI to identify intentions or emotions behind the words used

4 The ability of AI to identify specific types of object within an image

5 The ability of AI to comprehend the content of an image

6 Recurrent neural networks process data which has an ordered sequence to identify patterns and relationships within the series

7 https://research.nvidia.com/publication/2017-10_Progressive-Growing-of

8 https://arxiv.org/pdf/1703.10593.pdf

9 The availability of AI-skilled experts is one of the key areas covered by recommendations made by the UK Government's report "Growing the artificial intelligence industry in the UK", October 2017. https://www.gov.uk/government/publications/growing-the-artificial-intelligence-industry-in-the-uk

10 Examples of organisations offering these services include Crisp Thinking, Two Hat Security and Besedo

11 For example, the initial Facebook social networking service, then known as FaceMash, started in July 2003, but Facebook did not become cash-flow positive until it had 300 million users in September 2009. https://www.cbc.ca/news/technology/facebook-cash-flow-positive-signs-300m-users-1.826223

12 For example, YouTube has been criticised by advertisers for placing their adverts alongside violent videos made by Islamic State. https://www.economist.com/briefing/2019/05/04/the-tricky-task-of-policing-youtube

13 There have been some strong headlines in the press which have brought significant attention on the ethics of social media businesses, such as: "Instagram 'helped kill my daughter' ", BBC News, 22nd January 2019. https://www.bbc.co.uk/news/av/uk-46966009/instagram-helped-kill-my-daughter

14 John Suler identified the disinhibition effect in his 2004 paper in CyberPsychology and Behavior: https://www.liebertpub.com/doi/abs/10.1089/1094931041291295

15 http://www.kevinmunger.com/

16 https://www.gov.uk/government/publications/growing-the-artificial-intelligence-industry-in-the-uk

17 https://www.ofcom.org.uk/__data/assets/pdf_file/0022/120991/Addressing-harmful-online-content.pdf

18 https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/650949/Internet_Safety_Strategy_green_paper.pdf

19 https://www.gov.uk/government/consultations/internet-safety-strategy-green-paper

20 http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

21 From Merriam-Webster dictionary: https://www.merriam-webster.com/dictionary/artificial%20intelligence

22 The Turing Test is a hypothetical test in which an interrogator tries to differentiate between a human and a machine, based purely on their written responses to written questions. The test is frequently used as a measure of artificial intelligence; however, the test does not explicitly measure intelligence, only the ability of a machine to imitate human intelligence.

23 https://books.google.co.uk/books?id=2FMEAAAAMBAJ&pg=PA57#v=onepage&q&f=false

24 https://www.theclever.com/15-huge-supercomputers-that-were-less-powerful-than-your-smartphone/

25 https://scryanalytics.ai/genesis-of-ai-the-first-hype-cycle/

26 https://www.dataversity.net/brief-history-artificial-intelligence/

27 https://twitter.com/fchollet/status/940992948786806784

28 https://www.forbes.com/sites/louiscolumbus/2017/10/18/cloud-computing-market-projected-to-reach-411b-by-2020/#b531dd578f29

29 https://cloud.google.com/blog/products/gcp/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu

30 Video of progress during learning: https://www.youtube.com/watch?v=cAPPg_r18ec

31 https://www.forbes.com/sites/quora/2016/08/05/this-is-the-cutting-edge-of-deep-learning-research/#7df14eaa51c8

32 https://www.gov.uk/government/publications/growing-the-artificial-intelligence-industry-in-the-uk/recommendations-of-the-review

33 https://www.gov.uk/government/publications/artificial-intelligence-sector-deal/ai-sector-deal

34 https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-applications-and-value-of-deep-learning

35 https://www.aspect.com/globalassets/2016-aspect-consumer-experience-index-survey_index-results-final.pdf

36 https://www.businesswire.com/news/home/20180703005029/en/Juniper-Research-Chatbots-Deliver-11bn-Annual-Cost

37 The Society of Automotive Engineers have defined six levels of autonomous driving capability. These range from level 0, in which autonomous systems issue warnings and may momentarily take control, through to level 5, in which no human intervention is required at all. Most modern cars include features which provide some level 1 autonomy, such as adaptive cruise control or lane keeping assistance, but in level 1 the driver must be ready to retake full control at any time. https://www.sae.org/standards/content/j3016_201806/

38 https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology

39 https://www.tensorflow.org/

40 https://opensource.google.com/projects/deepmind-lab

41 https://facebook.ai/developers/tools

42 https://blog.openai.com/universe/

43 https://www.ft.com/content/428f1502-2b5c-11e8-9b4b-bc4b9f08f381

44 https://www.seas.harvard.edu/news/2017/06/harvard-launches-data-science-master-s-degree-program

45 https://www.intel.co.uk/content/www/uk/en/internet-of-things/infographics/guide-to-iot.html

46 http://www.image-net.org/about-overview

47 https://www.gartner.com/en/newsroom/press-releases/2018-09-12-gartner-forecasts-worldwide-public-cloud-revenue-to-grow-17-percent-in-2019

48 https://www.forbes.com/sites/louiscolumbus/2018/09/23/roundup-of-cloud-computing-forecasts-and-market-estimates-2018/#128915fd507b

49 https://www.ft.com/content/1a0ed6c8-18ee-11e9-b93e-f4351a53f1c3

50 https://arxiv.org/pdf/1812.02391.pdf

51 https://arxiv.org/pdf/1707.07012.pdf

52 https://towardsdatascience.com/everything-you-need-to-know-about-automl-and-neural-architecture-search-8db1863682bf

53 https://www.technologyreview.com/s/610421/on-device-processing-and-ai-go-hand-in-hand/

54 https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

55 https://www.research.ibm.com/artificial-intelligence/publications/?researcharea=explainability

56 https://www.darpa.mil/program/explainable-artificial-intelligence

57 http://moreisdifferent.com/2017/09/hinton-whats-wrong-with-CNNs

58 https://www.etsi.org/technologies/multi-access-edge-computing

59 https://www.itu.int/en/ITU-T/focusgroups/ml5g/Pages/default.aspx

60 Such as "Unified architecture for machine learning in 5G and future networks", January 2019, https://www.itu.int/en/ITU-T/focusgroups/ml5g/Documents/ML5G-delievrables.pdf

61 Most recently in August 2018: https://www.itu.int/en/ITU-T/Workshops-and-Seminars/20180807/Pages/default.aspx

62 Nokia, August 2018: https://www.itu.int/en/ITU-T/Workshops-and-Seminars/20180807/Documents/Ran_ML.pdf

63 https://www.businessinsider.com/netflix-recommendation-engine-worth-1-billion-per-year-2016-6?r=US&IR=T

64 https://medium.com/netflix-techblog/artwork-personalization-c589f074ad76

65 https://www.bbc.co.uk/news/newsbeat-45939044

66 https://news.developer.nvidia.com/ai-can-convert-black-and-white-clips-into-color/

67 https://blog.openai.com/better-language-models/

68 https://www.accenture.com/us-en/new-delivery-reality-post-parcel-players-index

69 https://insideretail.asia/2017/06/07/how-alibaba-uses-artificial-intelligence-to-change-the-way-we-shop/

70 https://www.forbes.com/sites/insights-intelai/2018/09/21/beyond-connectivity-three-strategies-for-telecom-growth/

71 https://hbr.org/2017/06/how-ai-is-streamlining-marketing-and-sales

72 http://www.telcoprofessionals.com/blogs/30408/1213/will-ai-end-fraud-in-telecoms

73 https://www.ft.com/content/0dca8946-05c8-11e8-9e12-af73e8db3c71

74 https://support.discordapp.com/hc/en-us/articles/115000982752-Screen-sharing-Video-Calls

75 https://www.bbc.co.uk/news/business-47620519

76 https://www.theguardian.com/technology/2017/apr/17/facebook-live-murder-crime-policy

77 https://econsultancy.com/memes-in-marketing-seven-memorable-examples-from-brands/

78 https://code.fb.com/ai-research/rosetta-understanding-text-in-images-and-videos-with-machine-learning/

79 https://www.bbc.co.uk/news/technology-42912529

80 https://www.bath.ac.uk/announcements/ai-could-make-dodgy-lip-sync-dubbing-a-thing-of-the-past/

81 https://transparencyreport.google.com/youtube-policy/removals?content_by_flag=period:Y2018Q3;exclude_automated:&lu=content_by_flag

82 https://www.theguardian.com/world/2017/jan/02/facebook-blocks-nude-neptune-statue-bologna-italy

83 https://money.cnn.com/2016/07/07/media/facebook-live-streaming-police-shooting/

84 https://transparencyreport.google.com/youtube-policy/removals

85 https://www.businessinsider.com/facebook-releases-data-content-moderation-2018-5?r=UK

86 https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html

87 https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/

88 https://www.theguardian.com/technology/2016/sep/09/facebook-reinstates-napalm-girl-photo

89 https://knowyourmeme.com/memes/triple-parentheses-echo

90 https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine

91 https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/787736/CDEI_2_Year_Strategy.pdf

92 https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

93 https://harvardlawreview.org/wp-content/uploads/2018/04/1598-1670_Online.pdf

94 https://www.wired.com/2014/10/content-moderation/

95 https://www.imdb.com/title/tt7689936/

96 https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

97 http://www.technologycoalition.org/wp-content/uploads/2012/11/EmployeeResilienceGuidebookFinal7-13.pdf

98 https://arstechnica.com/tech-policy/2013/04/wikipedia-editor-allegedly-forced-by-french-intelligence-to-delete-classified-entry/

99 https://newsroom.fb.com/news/2018/01/news-feed-fyi-bringing-people-closer-together/

100 https://www.bbc.co.uk/news/blogs-trending-39381889

101 https://arxiv.org/pdf/1602.04938.pdf

102 https://www.darpa.mil/attachments/XAIProgramUpdate.pdf

103 https://arxiv.org/pdf/1612.07360.pdf

104 https://arxiv.org/pdf/1806.07421.pdf

105 https://www.darpa.mil/attachments/XAIProgramUpdate.pdf

106 https://arxiv.org/pdf/1704.05796.pdf

107 https://www.bbc.co.uk/news/business-34324772

108 http://www.internetlivestats.com/twitter-statistics/

109 https://gifct.org/about/

110 https://www.microsoft.com/en-us/PhotoDNA/

111 https://support.google.com/youtube/answer/2797370?hl=en-GB

112 https://www.bbc.co.uk/news/technology-47278362

113 https://newsroom.fb.com/news/2018/05/enforcement-numbers/

114 https://arxiv.org/pdf/1301.3781.pdf

115 https://arxiv.org/pdf/1706.03762.pdf

116 https://arxiv.org/pdf/1708.00524.pdf

117 https://affect.media.mit.edu/pdfs/IJCAI-DinakarPicardLieberman(1).pdf

118 https://code.fb.com/ai-research/unsupervised-machine-translation-a-novel-approach-to-provide-fast-accurate-translations-for-more-languages/

119 https://www.microsoft.com/en-us/research/blog/microsoft-researchers-algorithm-sets-imagenet-challenge-milestone/

120 http://web.media.mit.edu/~kdinakar/a18-dinakar.pdf

121 https://pdfs.semanticscholar.org/100b/09552f77abba945a297cbbb1dce8ee3c986e.pdf

122 Cybersafety2016.github.io/slides/Cybersafety2016_keynote.pdf

123 https://blog.twitter.com/official/en_us/topics/company/2018/how-twitter-is-fighting-spam-and-malicious-automation.html

124 https://academic.oup.com/bjc/article/56/2/211/2462519

125 https://www.sciencedirect.com/science/article/pii/S0957417417300751

126 https://research.nvidia.com/publication/2017-10_Progressive-Growing-of

127 https://blog.openai.com/better-language-models/

128 https://danklearning.com/

129 https://arxiv.org/pdf/1806.04510.pdf

130 https://arxiv.org/pdf/1703.10593.pdf

131 https://www.researchgate.net/publication/319672232_Effective_data_generation_for_imbalanced_learning_using_Conditional_Generative_Adversarial_Networks

132 https://code.fb.com/core-data/introducing-deeptext-facebook-s-text-understanding-engine/

133 https://code.fb.com/ml-applications/under-the-hood-suicide-prevention-tools-powered-by-ai/

134 https://www.investopedia.com/articles/active-trading/111115/why-all-worlds-top-10-companies-are-american.asp

135 For example: BBC News: "Instagram 'helped kill my daughter' ", 22nd January 2019, https://www.bbc.co.uk/news/av/uk-46966009/instagram-helped-kill-my-daughter

136 For example: Mark Zuckerberg, the founder of Facebook attended a hearing with the US Congress in early April 2018

137 For example: House of Commons Select Committee: "Mark Zuckerberg 'not able' to attend unprecedented international joint hearing in London", 14th November 2018 https://www.parliament.uk/business/committees/committees-a-z/commons-select/digital-culture-media-and-sport-committee/news/facebook-letter3-17-19/

138 https://harvardlawreview.org/wp-content/uploads/2018/04/1598-1670_Online.pdf

139 https://www.gov.uk/government/publications/growing-the-artificial-intelligence-industry-in-the-uk

140 https://aws.amazon.com/rekognition/

141 https://cloud.google.com/vision/

142 https://azure.microsoft.com/en-gb/services/cognitive-services/content-moderator/

143 https://xmoderator.com/

144 https://www.clarifai.com/

145 https://www.crispthinking.com/

146 https://www.twohat.com/

147 https://besedo.com/

148 https://towardsdatascience.com/comparison-of-the-best-nsfw-image-moderation-apis-2018-84be8da65303

149 CyberPsychology and Behavior, John Suler, June 2004 available at: https://www.liebertpub.com/doi/abs/10.1089/1094931041291295

150 https://www.theguardian.com/technology/2017/mar/21/youtube-google-advertising-policies-controversial-content

151 https://www.polygon.com/2014/3/20/5529784/how-riot-games-encourages-sportsmanship-in-league-of-legends

152 https://newsroom.fb.com/news/2018/01/news-feed-fyi-bringing-people-closer-together/

153 https://www.theguardian.com/technology/2018/may/15/twitter-ranking-algorithm-change-trolling-harassment-abuse

154 https://www.bbc.co.uk/news/uk-46976753

155 https://www.huffingtonpost.com/shane-paul-neil/more-than-half-of-america_b_7872514.html?ec_carp=3182375576800502504

156 https://arxiv.org/pdf/1504.00680.pdf

157 http://aclweb.org/anthology/P18-2031

158 https://www.ibm.com/blogs/research/2018/07/offensive-language-social-media/

159 https://www.ibm.com/blogs/research/2018/07/offensive-language-social-media/

160 https://arxiv.org/pdf/1504.02305v1.pdf

161 https://www.sciencemag.org/news/2017/05/bad-bots-do-good-random-artificial-intelligence-helps-people-coordinate

162 http://aclweb.org/anthology/P18-2031

163 https://www.washingtonpost.com/news/monkey-cage/wp/2016/11/17/this-researcher-programmed-bots-to-fight-racism-on-twitter-it-worked/

164 http://eprints.lse.ac.uk/37810/1/Designing_for_nudge_effects_%28lsero%29.pdf

165 http://www.kevinmunger.com/

166 https://www.gov.uk/government/publications/growing-the-artificial-intelligence-industry-in-the-uk

For more information, please contact:

Tim Winchcomb, Head of Technology Strategy – Wireless and Digital Services
tim.winchcomb@cambridgeconsultants.com

About Cambridge Consultants

Cambridge Consultants is a world-class supplier of innovative product development, engineering and technology consulting. We work with organisations globally to help them manage the business impact of the changing technology landscape.

With a team of more than 850 staff in the UK, the USA, Singapore and Japan, we have all the in-house skills needed to help you – from identifying the impact of emerging technologies on markets, to creating innovative concepts right through to taking your product into manufacturing. Most of our projects deliver prototype hardware or software and trial production batches. Our technology consultants can help you to develop strategies, policies and plans that take account of the step change often delivered by cutting-edge technologies.

We're not content to create 'me-too' solutions that make incremental change: we specialise in helping companies achieve the seemingly impossible. We work with some of the world's largest blue-chip companies as well as with some of the smallest, most innovative start-ups who want to change the status quo fast.

Cambridge Consultants is part of the Altran Group, a global leader in innovation. www.Altran.com

Cambridge Consultants

UK ▪ USA ▪ SINGAPORE ▪ JAPAN

www.CambridgeConsultants.com

Cambridge Consultants is part of the Altran group, a global leader in innovation. www.Altran.com