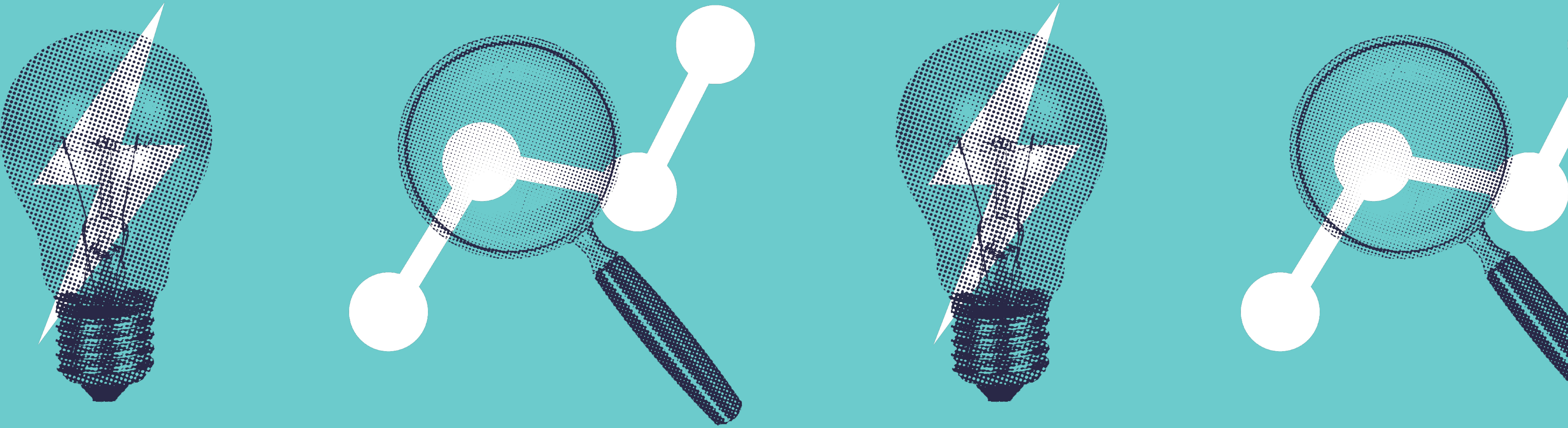
The role of AI in addressing misinformation on social media platforms

Centre for
Data Ethics
and Innovation



About the CDEI

The Centre for Data Ethics and Innovation (CDEI) is a government expert body that enables trustworthy use of data and AI. Our multidisciplinary team of specialists are supported by an advisory board to deliver, test and refine trustworthy approaches to data and AI governance, working in partnership with public sector and industry bodies.

Our goal is to create the conditions in which trustworthy innovation can thrive: an environment in which the public are confident their values are reflected in the way data-driven technology is developed and deployed; where we can trust that decisions informed by algorithms are fair; and where risks posed by innovation are identified and addressed.

For more information about the discussion or the CDEI's work more broadly, please get in touch at cdei@cdei.gov.uk.

Centre for
Data Ethics
and Innovation

Introduction

- Over the past year, misinformation has consistently undermined the global public health response to the COVID-19 pandemic. It has led to the use of dangerous and false health cures, increased the spread of the virus by perpetuating a myth that it is fake or presents no risk, and slowed vaccine uptake. All of this has contributed, directly or indirectly, to the deaths of many around the world, while also serving to deepen the public's distrust in democratic institutions.

- In a bid to reduce the volume of COVID-19 misinformation, platforms have over the last year introduced a number of new measures and policies, many of which rely on the use of algorithms to help detect and address harmful content. At the same time, platforms have been forced to turn to machine-based content moderation in order to cover for shortfalls in their workforce of human moderators, many of whom were unable to work from the office for prolonged periods.

- While initially proposed as a temporary solution to a unique crisis, some have questioned whether automated content moderation could become part of the status quo. Advocates see little choice but to rely heavily on algorithms to rein in misinformation. Critics, however, say that algorithms are not equipped to identify harmful content, and believe their adoption at scale could lead to unintended censorship. Others see this debate as a distraction from the more important question of how to reform platform business models, which may perpetuate misinformation.

Aims

- To help shed light on these matters, the CDEI hosted an expert forum that brought together a range of stakeholders, including platforms, fact-checking organisations, media groups, and academics. We sought to understand:
    - **The role of algorithms in addressing misinformation on platforms**, including what changed during the pandemic and why, and the limits of what algorithms can do;
    - How much platforms tell us about the role of algorithms within the content moderation process, and **the extent to which there should be greater transparency** in this regard;
    - Views on **the effectiveness of platform approaches to addressing misinformation**, including where there may be room for improvement in the immediate future.
- Debates relating to content moderation and misinformation cut across multiple issues, from the philosophical (e.g. where the right to freedom of speech begins and ends) to the cultural (e.g. how we define what is harmful and offensive). With limited time at our disposal we were unable to cover every area. Instead we focused on the questions above, which largely relate to the technical efficacy of content moderation tools and policies. We do note, however, where normative questions present particular challenges for addressing misinformation, and reference noteworthy work and developments in this field.

Recent and relevant work on misinformation

Our forum took place against the backdrop of a changing policy landscape. Recent and important interventions and investigations include the:

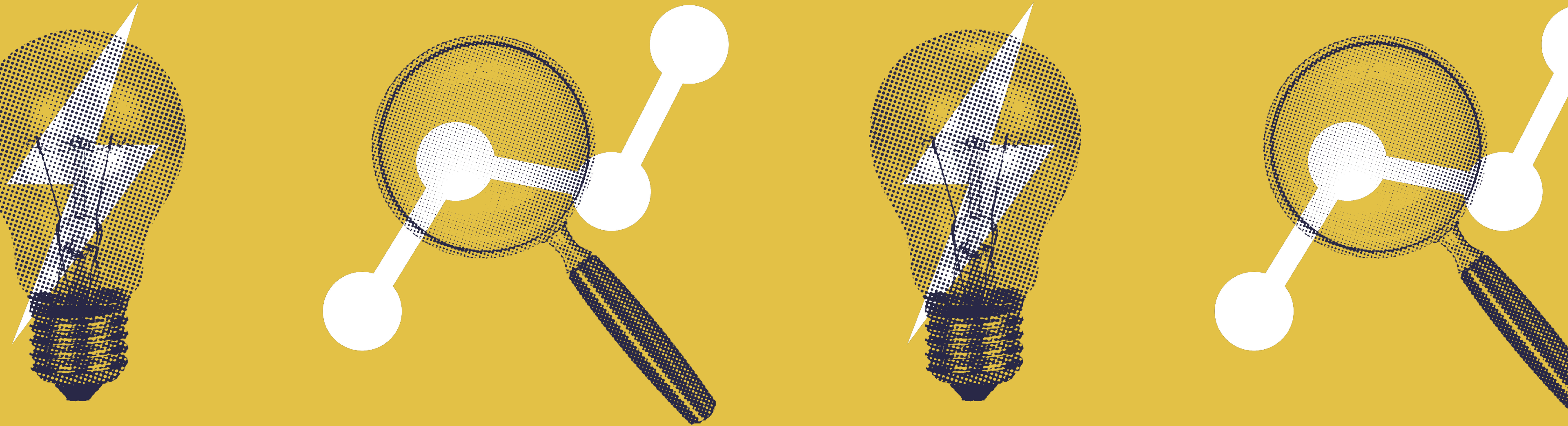
- Draft Online Safety Bill **(DCMS):** This establishes a new regulatory framework to tackle harmful content online, with Ofcom taking on the role of regulator. The Bill will establish new duties of care for online service providers in relation to certain harmful content, which may include misinformation.
- Review of online targeting **(CDEI):** This looked at the impact of content recommendation systems on platforms and what measures might be introduced to make them more accountable and transparent, and to empower users. We highlighted the need to strengthen regulatory oversight of online targeting systems through the proposed online harms regulator (Ofcom), looking in particular at how to protect people's freedom of expression and privacy.
- Online Safety Data Initiative **(DCMS):** This aims to facilitate better access to data relating to online harms, working with technology companies, service providers, civil society and academics, as well as the CDEI.

Participants

- Imran Ahmed, Centre for Countering Digital Hate
- Charlie Beckett, LSE
- Ellen Judson, Demos
- Lisa-Maria Neudert, Oxford Internet Institute
- Will Moy, Full Fact
- Matt Rogerson, Guardian
- Myrna MacGregor, BBC
- Emre Kizilkaya, International Press Institute
- Richard Earley, Facebook
- Alina Dimofte, YouTube
- Elizabeth Kanter, TikTok
- Sally Lehrman, The Trust Project
- Emma Haselhurst, Logically
- Dhruv Ghulati, Factmata

Key findings

Centre for
Data Ethics
and Innovation



Key findings

- **Algorithms play an essential role in moderating content on social media platforms.** They can be used to identify material that has already been banned, preventing the upload of images, text and videos that are known to be harmful. They can also help to detect previously unseen forms of misinformation, by identifying signals that are indicative of malicious content. These tools, some of which are based on machine learning, provide the capacity to manage content at a speed and scale that would not be possible for human moderators operating alone.

- Prior to the pandemic, algorithms would have typically been used in conjunction with humans, helping to identify and categorise content to inform the decisions of trained staff. However, **the onset of COVID-19 and resulting lockdown led to a reduction in the moderation workforce**, just as the volume of misinformation was rising. Platforms responded by relying on automated content decisions to a greater extent, without significant human oversight.

- The platforms taking part in our forum acknowledged that **this increased reliance on algorithms led to substantially more content being incorrectly identified as misinformation.** Participants noted that algorithms still fall far short of the capabilities of human moderators in distinguishing between harmful and benign content. One reason is that misinformation is often subtle and context dependent, making it difficult for automated systems to analyse. This is particularly true for misinformation that relates to new phenomena - such as COVID-19.

- Platforms have issued reassurances that the increased reliance on algorithms is only temporary, and that human moderators will continue to be at the core of their processes. However, it does appear that some companies have yet to revert to their pre-pandemic practices. **Platforms could do more to explain what changes are being made and how long they are expected to be in place.**

Key findings

- Content moderation, however, requires more than the detection of harmful misinformation. It also involves finding the right way to present that content to users - a policy decision enacted by a technical system. Where content is deemed illegal or extremely harmful, it may be removed from a platform. Yet **in many cases, misinformation content is not removed but presented in a way that changes how the user engages with it.** Platforms can downrank content, apply fact-checking labels, and promote truthful and authoritative information.

- The platforms represented at our forum highlighted examples of where these measures have had a clear and positive impact on how users engage with misinformation. However, some participants were sceptical about their effectiveness, believing that nothing short of removing harmful content would suffice. Some participants also argued that the commercial models of platforms, which are driven by user engagement and interactions, are **fundamentally at odds with addressing misinformation.**

- **A lack of evidence may be hindering our understanding of "what works"** in the moderation of content. Platforms say they have taken steps to support independent research on the aforementioned methods, including by providing researchers with access to their internal data. Yet participants at our forum believed that more could be done. Further research could, for example, help to shed light on whether the efficacy of measures varies by demographic group or geographic region. **Without such evidence, it will remain difficult to meaningfully scrutinise platform behaviour.**

- Participants felt that platforms could also be more transparent about their policies, including how they use algorithms. Platforms acknowledge this, and have begun to disclose more information about how they deal with harmful content, for example via transparency reports that log data on content and account removals. In future, **platforms could go further, including by disclosing information about the precision of their algorithms, and by clarifying how they define "borderline" and rule-breaking content.**

Key findings

- Platforms emphasised the importance of having **clear guidance from the government** on the types of information they should be disclosing, how often and to whom**.** As the new online harms regulator, **Ofcom is well positioned to set new benchmarks for clear and consistent transparency reporting.**

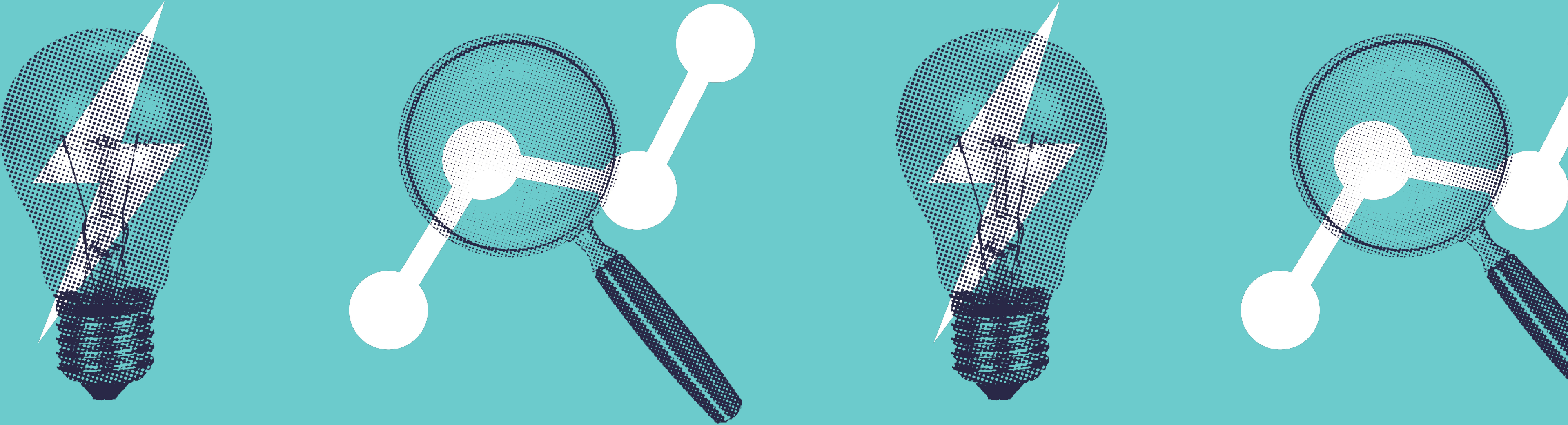
- While much of our discussion focused on the UK context, several participants expressed concern that **content moderation is even less effective in other parts of the world, particularly low income countries.** Many places lack an impartial media and strong civil society organisations that would otherwise rebut misinformation and be a source of truth for citizens. Some participants felt that platforms had an insufficient understanding of the political and cultural contexts of these countries, and that their algorithms were less effective in analysing non-Western languages. Greater investment in technological and human resources may be required to mitigate these risks.

- The forum went on to consider how platforms promote trustworthy media. **Official sources are prioritised by platforms, yet they still struggle to compete with misinformation** in terms of user engagement. Participants also said there was disagreement over what constitutes an "authoritative" source, with some feeling that platforms could be better at judging what is genuinely trustworthy, and doing more to support those outlets.

- **Altogether, participants were pessimistic about our collective capacity to resolve the challenges of misinformation in the immediate future.** Protecting truth on the internet will require shifts in our culture, technology and business practices, which will take time to realise. However, there are steps that we can take today to help mitigate misinformation. Undertaking more research into the efficacy of moderation tools, experimenting with new moderation methods, increasing transparency of platform moderation policies, and investing more in supporting authoritative content - all are interventions worthy of investigation.

Platforms and the spread of misinformation

Centre for
Data Ethics
and Innovation



What is misinformation?

- **Misinformation and disinformation are often used as interchangeable terms, but many experts define them as two distinct concepts** distinguished by the intention of the person disseminating the information**.** The UK government uses the following definitions**:**
  - **Disinformation** is the deliberate creation and dissemination of false and/or manipulated information that is intended to deceive and mislead audiences, either for the purposes of causing harm, or for political, personal or financial gain.
  - **Misinformation** refers to inadvertently spreading false information.
- The measures to address misinformation and disinformation online are very different, including the role of AI. To avoid an overly broad discussion and add to ongoing debates, we focused our work solely on measures to address misinformation.

LIVE Global coronavirus infections outpace China cases

Both the US and Europe are tightening restrictions further in a bid to halt the outbreak.

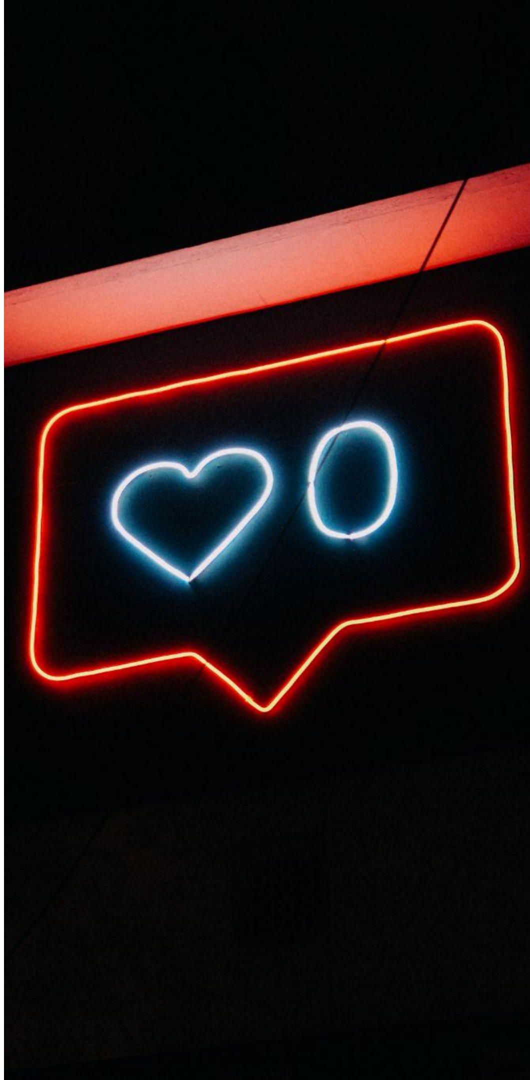
...cks fall despite stimulus

What is misinformation?

- **Misinformation has always existed and is not an entirely 'new' problem.** False rumours, incorrect reporting and conspiracy theories have been constant and inevitable throughout history. **The scale of harm arising from misinformation has varied,** ranging from causing many deaths (e.g. the AIDS epidemic) to being entirely innocuous (e.g. the search for the Loch Ness Monster).

- While misinformation has always been commonplace online, **the prevalence and impact of misinformation has increased substantially in recent years.** Previously fringe ideas and beliefs have become more widespread and mainstream, for example the QAnon and Pizzagate conspiracy theories. In recent years, platforms have also played host to significant amounts of political misinformation, which has undermined election integrity, sowed distrust and incited violence.

- Many critics blame online platforms in particular for failing to curtail the spread of misinformation promoted on their sites, and for doing too little, too late.



How do platforms risk increasing misinformation?

- **Platforms rely on engaging content to maintain and increase their user bases.** To achieve this, platforms such as Facebook, TikTok and YouTube focus on providing a personalised and relevant experience for their users.

- **Recommendation systems** are algorithms that take the data that platforms hold and use it to show content that users are likely to be interested in. This also provides more data about the user's interests and preferences to inform targeted advertising, which is crucial as advertising revenue is a key revenue stream for most platforms.

- Although recommendation systems can provide users with highly relevant content and personalised experiences, **they can easily promote harmful content such as misinformation.** Many forms of misinformation spread particularly rapidly because they are so engaging, often taking the form of sensational stories or entertaining memes, which may be shared by users and amplified by recommendation systems.

- As a result, platforms have faced criticism that they have **acted as catalysts for the spread of misinformation, while doing too little to curtail the negative impacts** of content recommendation systems.

What do platforms do to address misinformation?

- **All platforms undertake content moderation to detect and address harmful content, including misinformation.** Measures include:
  - **Removing content**
  - **Reducing the prominence of misinformation** in searches and user feeds, known as **downranking** content
  - **Labelling content** to indicate to users that it may be false, misleading, or some other form of misinformation
  - **Promoting authoritative sources of information** by making their content more prominent in searches and user feeds
  - **Increasing friction** in the user experience, usually to encourage users to take more time to think critically about the content they engage with and share
- Misinformation is generally a **legal harm**, meaning that platforms can **choose whether and how to address it,** unlike illegal harms such as extremist content which must be removed. **This results in a range of policies and approaches to addressing misinformation,** reflecting the differences in terms of design, users and the prevalence of misinformation across platforms.

What do platforms do to address misinformation?

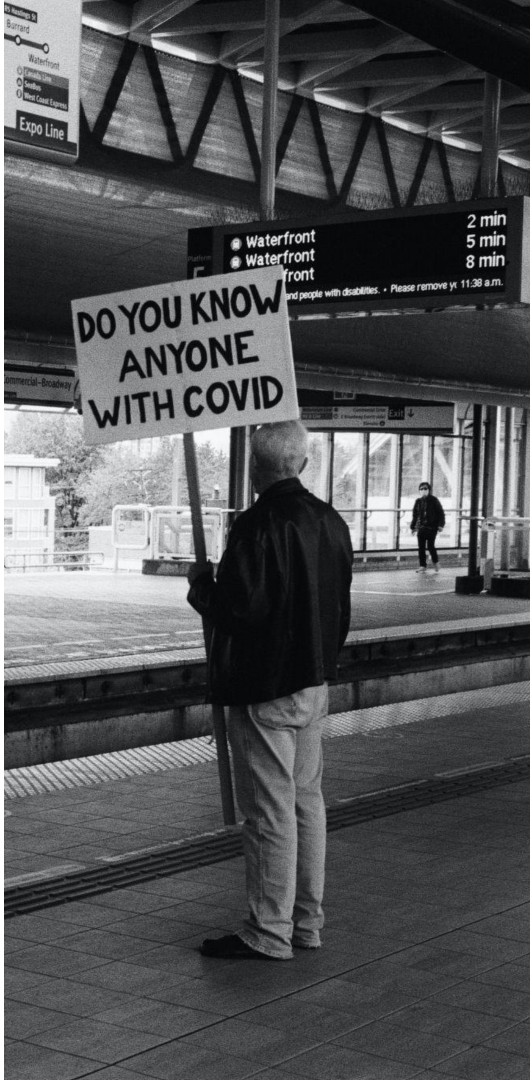
- Most large platforms allow a broad plurality of content, including views that may be offensive, objectively wrong, or even carry some risk of harm. **Platforms are keen to avoid being seen to over-police content**, or to appear as though they are restricting open debate. Yet **they also face pressure to do more to address misinformation online.** Being seen to censor content or failing to address misinformation could impact platforms' reputation with users, and by extension advertisers.

- While many platforms are keen to avoid being seen as the 'arbiters of truth', **since the start of the pandemic platforms have increasingly taken action to address misinformation**, expanding their policies on harmful misinformation. Many critics argue that platforms have been far too slow to develop misinformation policies, and the platforms we spoke to recognised this.

- **Platforms face a significant challenge in establishing the appropriate threshold for harmful misinformation that warrants moderation. The scale of harm arising from misinformation can vary substantially,** and while distinctions between harmful and harmless misinformation can seem straightforward at first glance, in practice this can be very difficult.

What do platforms do to address misinformation?

- **The platforms we spoke to noted the substantial difficulty in defining and categorising misinformation or levels of harm that might be presented by content.** While it can be easy to point to markedly different examples of misinformation, in practice a wide range of factors can influence the potential impact and harm arising from misinformation. This includes:
    - **The audience for the content**, for example whether vulnerable people are more likely to be exposed.
    - **The forum in which content is shared**, for example whether there is reliable information available to users that can help to counter false narratives.
    - **How widely and quickly the content can be spread.** This can be hard to determine, particularly for new forms of harmful misinformation, and will be dependent on audience and forum, as well as how content might be promoted by content recommendation systems.



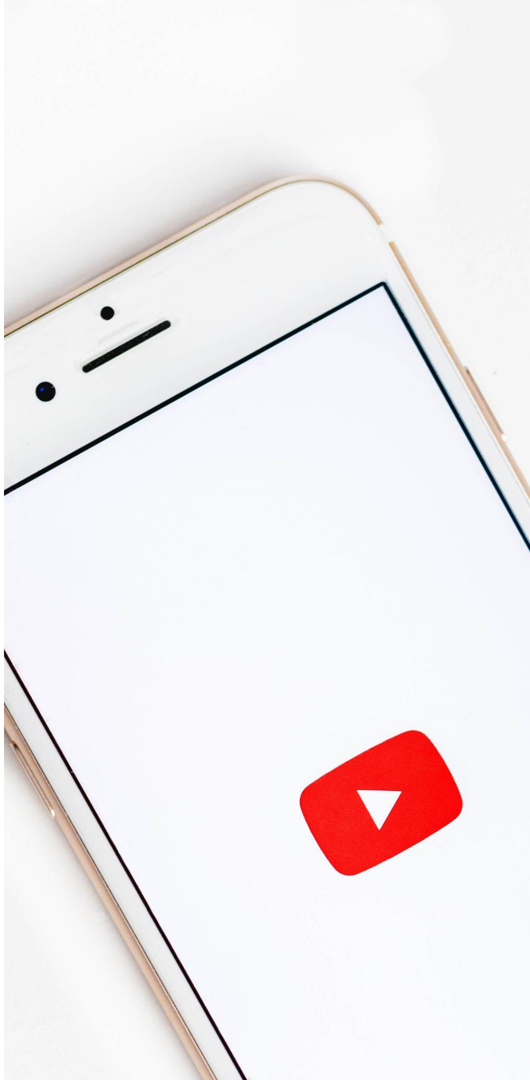
What role do algorithms play?

Platforms face significant challenges in managing the sheer volume of content posted by users each day. Detecting and taking action to address harmful misinformation on such a large scale **would be impossible by purely human means**. Platforms employ various algorithms to detect misinformation rapidly and at scale, making them **essential to the content moderation process.** This takes three forms:

- **Filtering:** the first line of defence for platforms before content goes live, algorithms are used to identify banned content such as videos and images and prevent it from being posted. Filtering only works for content that has already been seen before and flagged as rule-breaking, meaning that any new and 'unseen' harmful content will not be affected.

- **Detecting signs of misinformation:** algorithms are used to detect signals indicative of misinformation, flagging content for further review by a moderator. In recent years methods including natural language processing and computer vision have become substantially more sophisticated, allowing for more accurate detection across image, text, audio and video. These **algorithms provide scale and speed in flagging potential signs of misinformation. Facebook** has also begun to use algorithms to prioritise flagged content for review by a **human moderator.**

- **Content decisions:** increasingly algorithms have been used to automate content moderation decisions such as the removal of content once live on the platform, rather than leaving the decision to a human content moderator.

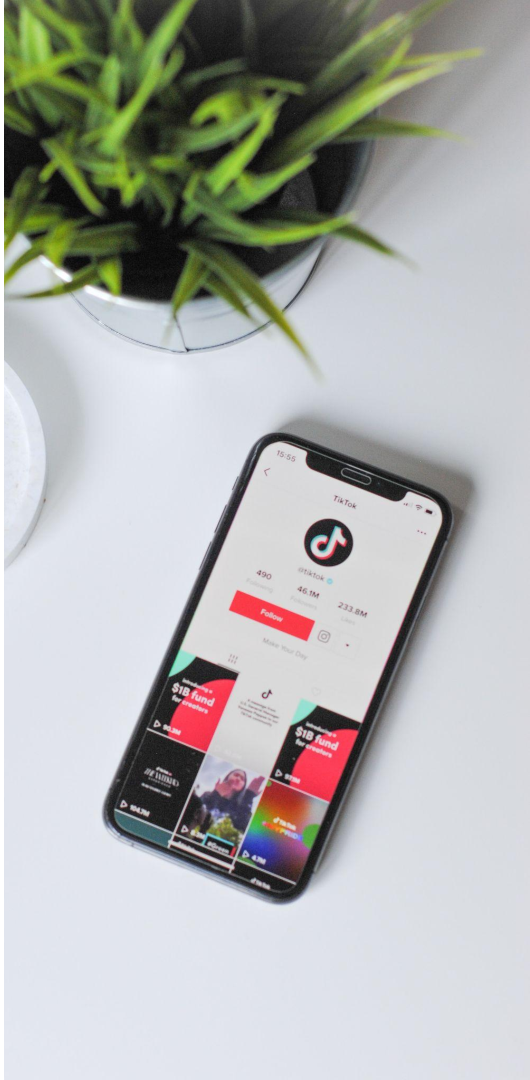


What role do algorithms play?

While algorithms are essential in addressing misinformation on platforms, they are not a cure-all. This is due to both the complex nature of misinformation and the current limits of algorithms for content detection and moderation.

- Platforms seek to identify and take action against as much rule-breaking content as possible (known as **true positives**) while minimising the amount of content that is wrongly identified as rule-breaking (known as **false positives**). With millions of posts uploaded every day on many platforms, it is inevitable that at least some rule-breaking content will be missed or wrongly classified as permissible (known as **true negatives**).

- **While algorithms can provide speed and scale in the content moderation process, they are also generally poor at contextual interpretation.** It may be easy for a human to understand that a post is satirical, or contains terrorist related content purely for educational purposes, but algorithms struggle with this nuance. **Misinformation is particularly challenging in this respect**, as it can be very subtle and context dependent, making it difficult even for humans to identify.



What role do algorithms play?

- **Different types of misinformation can also raise challenges for algorithmic content detection**. For example, participants in our discussion noted that it is **easier to train algorithms to identify 5G conspiracy theories than a broad range of false health cures**, as the former is narrower and may involve more consistent key words and phrases, while signals that distinguish false health cures from proven ones may be more difficult to establish. **New forms of misinformation are constantly emerging, which present an additional challenge** as algorithmic models are generally poor at responding to new information.

- These shortcomings were exemplified during the pandemic when many platforms had to rely on algorithmic methods to a greater extent. This is because, **at the start of the pandemic, Facebook, Twitter and YouTube had to reckon with a reduced and remote** content moderation workforce, while at the same time platforms have seen **increased user activity as people face local and national lockdowns.**

How did use of algorithms change during the pandemic?

- **Platforms increased their reliance on automated methods for content decisions, acknowledging that it would likely lead to an increase in false positives but that the alternative would be to risk more harmful content (true negatives) going undetected.**
  - **Facebook** decided to turn off the ability to request an official review of decisions due to capacity constraints, instead giving people the option to provide feedback. Based on this feedback they also saw an increase in reverted moderation decisions.
  - **While takedowns from YouTube doubled**, more than 50% of appeals were accepted and videos reinstated (compared to ~25% of appeals in normal times).
- This experience suggests platforms are unlikely to rely on fully automated content moderation in the immediate future. **Facebook, TikTok and YouTube all emphasised that human moderators will continue to remain at the core of their moderation processes.** While not discussed in our forum, many critics will likely question whether it is sufficient to revert to pre-pandemic practices, or whether additional measures, such as increased human moderation, are needed.

Are more human moderators needed?

The NYU Stern Center for Business and Human Rights' report, *Who Moderates the Giants? A Call to End Outsourcing*, highlights the challenges that many content moderators on many platforms face in resolving difficult content decisions with limited time to consider each post, all while facing the potential mental health impacts of reviewing extreme and highly graphic content.

The NYU report recommends doubling the number of content moderators to increase the time allocated for each content review, in order to increase the quality of decisions. It also suggests that moderators should be brought in-house, rather than being outsourced, to improve quality and benefits for moderators.

How did use of algorithms change during the pandemic?

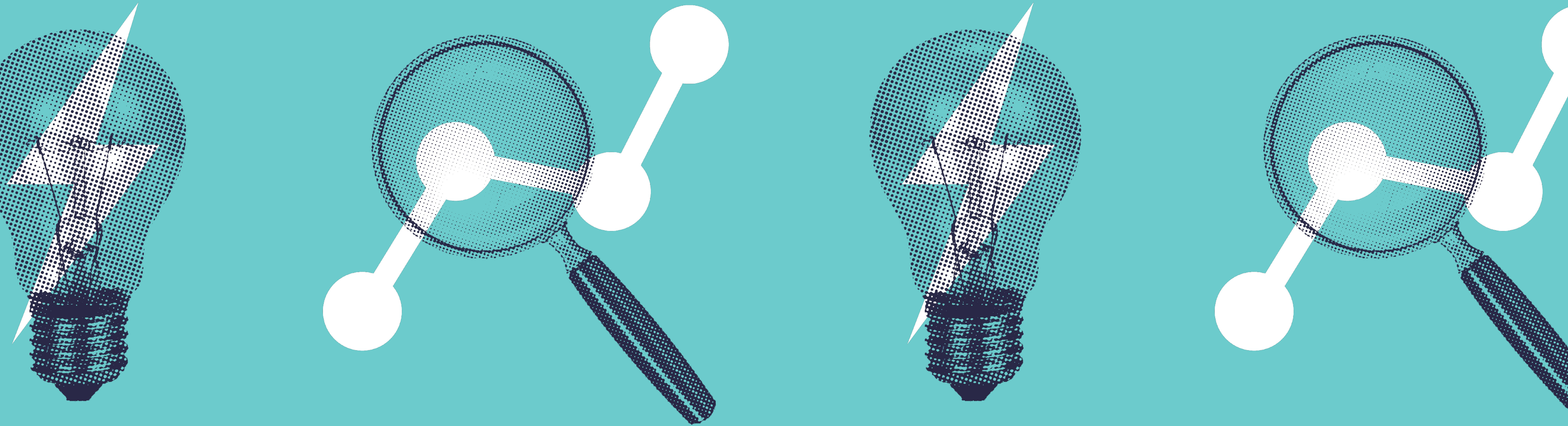
- Although platforms have emphasised that the increased reliance on automated content decisions is a temporary measure, **not all platforms have yet reverted to pre-pandemic practices.** For example, in April 2021 YouTube noted that it continues to rely on automated decisions to a greater extent in some regions due to the pandemic. **It is not clear when some of these practices will end,** precisely how the role of humans and algorithms has shifted over the past year, or how platforms intend to build learnings from this experience into the moderation process.

- As platforms continue to develop their use of algorithms in the content moderation process, **they should ensure that there is clarity regarding what changes are being made, and how this impacts the decision-making process**, particularly with respect to improvements such as precision and recall (increasing true positives and reducing false positives).

- **Platforms should also look to share their experiences of the challenges presented by moderation during the pandemic,** either through transparency reports or directly with regulators and trusted stakeholders where commercially sensitive information is involved.

Platforms measures to address misinformation

Centre for
Data Ethics
and Innovation



Pre-moderating content before it goes live

- In light of the limitations of content moderation, participants in our discussion **questioned why platforms do not pre-moderate content before it goes live on a platform**, as this could reduce the risk of harmful content being spread widely on a platform before being flagged and appropriately actioned. Pre-moderation would not prevent some harmful content going undetected by algorithms and remaining online.

- Pre-moderation would result in a **delay to content going live**, with the delay depending on the scale and thoroughness of the pre-moderation process. However, participants thought that some users would happily tolerate short delays if it led to a reduction in harmful content being posted (see right).

- Precedent from EU courts suggests that pre-moderation could be interpreted as 'generalised monitoring' of content, which **would result in greater platform liability for publishing content.** Such an interpretation is also possible in the UK, US and other parts of the world, and **could require a reshaping of the way platforms operate.** However, this could depend on the scale of pre-moderation, for example if it was limited to using algorithms to detect signals of only the most harmful content.

- Technological advances could make pre-moderation more feasible in the future by reducing the length of the posting delay, but **pre-moderation currently presents risks for platforms that make changes unlikely in the immediate future. It would be valuable to explore whether the law could or should be updated** to allow certain forms of pre-moderation without constituting generalised monitoring or editorial responsibility.

Quality Control

Demos' 2019 report *Quality Control* found that 59% of UK adults think that social media content should be edited by moderators, while 24% think that it should not. Younger people were less likely to think that content should be edited than older people, but younger people were more likely to have negative views about their use of social media.

Downranking misinformation

- Downranking content involves adjustments to platform recommendation systems in order to make certain content appear less prominently in feeds or searches while still being viewable by those who specifically seek it. This allows platforms to reduce engagement and interaction with misinformation without removing it.

- Some participants saw downranking as a **limited mitigation** of the impacts of content recommendation, with some arguing that it is simply not as effective as removing content or users who consistently post misinformation.

- Although the precise boundaries of what should be actioned and how is be context dependent, there was a view among some participants that **too little content is classed as rule-breaking or warranting removal,** while categorising content as 'borderline' creates an opaque standard for moderation decisions that may appear arbitrary.

- Equally, **users are generally not informed if their content is downranked,** which may lead them to think that their content is being hidden arbitrarily or for unclear purposes. **Providing explanations for these decisions,** as many platforms do for removal of content and fact-checking labels, **could inform users about how their content is being moderated and why.**

The Centre for Countering Digital Hate and removing high-profile accounts to prevent the spread of misinformation

In their report *The Disinformation Dozen*, the Centre for Countering Digital Hate highlights the leading voices in the anti-vaccination movement who play a key role in spreading misinformation on platforms, with 65% of anti-vaccination content on Facebook and Twitter 1 February 2021 - 16 March 2021 months being attributable to them. CCDH calls for platforms to remove these individuals entirely, arguing that they have consistently violated terms of service, and that deplatforming is the most effective means of reducing the spread of misinformation.

Some critics disagree that deplatforming is an effective approach, arguing that it can drive deplatformed actors and groups onto less visible and unregulated platforms, with the effect of further radicalising people.

Promoting authoritative data

- **A free and diverse press has been vital in dispelling misinformation during the pandemic**, particularly where audiences have low trust in governments or what they read on social media, or even where the government may actively promote misinformation. **Facebook** takes steps to promote informative media by prioritising sources that it has categorised as trusted and including a broader promotion of informative content.

- Many participants in our forum thought that **platforms could do more to support trustworthy media,** with concerns that platforms such as Facebook may prioritise western and national scale media over local sources, particularly in the developing world. Local media is crucial in informing communities during the pandemic, but **smaller outlets can struggle to gain trusted status from platforms,** and often do not have the resources to navigate platform recommendation systems or search engine optimisation effectively. Participants raised concerns about specific outlets being treated as trustworthy by Facebook, for example Breitbart.

- Although some are optimistic that promoting authoritative information is the best means of combating misinformation, many of our participants were sceptical. **Misinformation can spread rapidly, while the running of media organisations is costly and producing content is time consuming, making it difficult to compete with the scale and speed at which misinformation spreads online.**

The Trust Project and promoting trustworthy journalism

The Trust Project aims to amplify journalists' commitment to transparency, accuracy, inclusion and fairness so that the public can make informed news choices. Over 200 news sites have partnered with the Trust Project, implementing 8 Trust Indicators as a standard including (amongst other things) journalist expertise, local sourcing, ethical practices for news gathering and citations. The Trust Project partners with Facebook and Google to provide consistent standards for promoting reliable news.

Labels and fact-checking

Applying banners and labels to promote official information and advice

- Most platforms have also included official information across their home pages, as well as issuing tags and links to trustworthy guidance, such as how to register to vote or how to sign up for a COVID-19 vaccination. Platforms have used such tags and banners widely during the pandemic. For example, TikTok has claimed that its World Health Organisation (WHO) banners were at one point responsible for the majority of overall user traffic to the WHO website. However, platforms also acknowledged that while these measures can be useful, they can also see diminishing returns.

Fact-checking and applying labels to content

- **Where content is identified by a fact-checker as false or misleading, a platform may apply a label to make users aware that the information they interact with may be misleading, disputed or false**. Labelling allows content to remain online, but with caveats that can allow for scrutiny, open debate and education of users in order to ensure they are better equipped to recognise and respond to misinformation in the future. **TikTok** has found that measures to flag unsubstantiated videos to users and creators led to a 24% reduction in sharing such content, while likes reduced by 7%.
- While some critics fear that labels may backfire, and lead to distrust among some users, Full Fact has conducted research suggesting there is limited evidence for this argument. **Participants' greater concern was that labels may simply be ineffective,** as many users could ignore or simply fail to notice them, particularly as labels become more commonplace.

The challenges of online fact-checking

Full Fact's December 2020 report on fact-checking provides a detailed account of how fact checkers find, select and review content, incorporating experiences from around the world. Full Fact notes that Facebook's third party fact-checking programme should be commended, but improvements could be made including better information sharing with fact-checkers.

The report provides a number of recommendations for platforms, including that Google, YouTube and Twitter invest in paid fact-checking partnerships with third parties to improve impartiality and local knowledge.

Labels and fact-checking

- To combat this, platforms such as Facebook and Twitter have introduced fact-checking labels that hide the content of a post, requiring users to click on the post to see it. By adding additional steps to view or share content, platforms can significantly reduce the risk that users simply ignore fact-checking labels. Facebook claim this practice can reduce content views by up to 95%. **This is an example of increasing friction in the user experience to direct users away from certain actions**, measures that platforms have increasingly introduced over the past year.

- **It is difficult to assess the impact and value of measures to address misinformation with limited evidence for their effectiveness.** As we noted in our review of online targeting in relation to recommendation systems, **platforms could do more to ensure that independent researchers have** access to relevant data. **Platforms have acknowledged this, and TikTok is seeking to conduct research with external partners to better understand the impact of labels.** A growing body of external research is beginning to inform our understanding of the impact of labels and fact-checking.

- Platforms hold significant amounts of information about how users interact with content moderation decisions, as well as how changes to platform design alter user behaviour. Making more of this information available to trusted external stakeholders and the public could help inform our understanding of the effects of various measures to address misinformation. This is considered further in the following section.

Research into the impact of fact-checking labels

Research relating to the effects of fact-checking is still in its early stages, and has produced mixed findings. Zhang et al.'s research into the effects of fact-checking social media misinformation on attitudes towards vaccines suggests that labels posted immediately below misinformation can make users have more positive views about vaccines.

However Pennycook et al. have noted a potential unintended consequence of fact-checking misinformation, as it can lead viewers to believe other content by virtue of the fact that it has not been fact-checked, creating an 'implied truth' effect.

These examples underscore the need for much further research including working with platforms, as well as an awareness of the potential side-effects of certain approaches to labelling misinformation.

Labels and fact-checking

- Participants noted that **the greatest vulnerabilities to misinformation online may lie in countries with fewer resources and capabilities for content moderation.** Many countries have struggled with misinformation in social and traditional media while also suffering from weak democratic institutions and limited press freedoms.

- Participants noted that many countries lack the necessary civil society institutions as well as skills and resources for fact-checking and promoting trustworthy information. They also flagged that **platforms have fewer resources invested in certain regions and languages, resulting in a more limited local cultural understanding.**

- This may increase the risk that misinformation is not flagged or appropriately fact-checked, that new forms of misinformation are identified more slowly, and that platforms promote sources that are not trustworthy. An unsophisticated understanding of local contexts may also make it more difficult for trustworthy media to gain traction online.

- Additionally, research suggests that natural language processing techniques are less advanced in some languages, and so **may fail to detect signs of misinformation or produce more false positives. Greater investment in content moderation infrastructure could help reduce these risks,** although this may be challenging where there is limited local expertise, or more limited language datasets to train algorithms.

The IPI's Turkey Digital Media Report 2021

The Vienna-based International Press Institute's (IPI) *Turkey Digital Media Report 2021*, authored by Emre Kizilkaya and Burak Utucu, is an extensive study of online media in this country. The report includes data-based insights on how Google and YouTube algorithms distribute news content from Turkey's pro-government and independent outlets.

Closed networks

Throughout the pandemic, misinformation has spread rapidly via WhatsApp and other closed communication platforms. Closed networks present a number of challenges for dealing with misinformation:

- **It is very difficult to know the scale and speed at which misinformation spreads, because messages are generally not monitored.** It will often only become clear what content is being widely spread after some time.
- **Users may be less sceptical of the content they receive, as it is generally sent by friends or family.** The network effects of information are well known, and are also felt on open platforms.
- **In addition, there is no opportunity for moderation such as fact-checking or labelling content.** Many communication platforms are valued for the privacy that they provide to users, and so any changes could prompt users to find an alternative platform.
- **Some communities may be affected more than others.** Certain closed platforms may be popular amongst different age groups or communities, increasing the spread of misinformation. For example, the sharing of COVID-19 misinformation on WhatsApp in some orthodox Jewish communities in New York has been linked to local outbreaks.

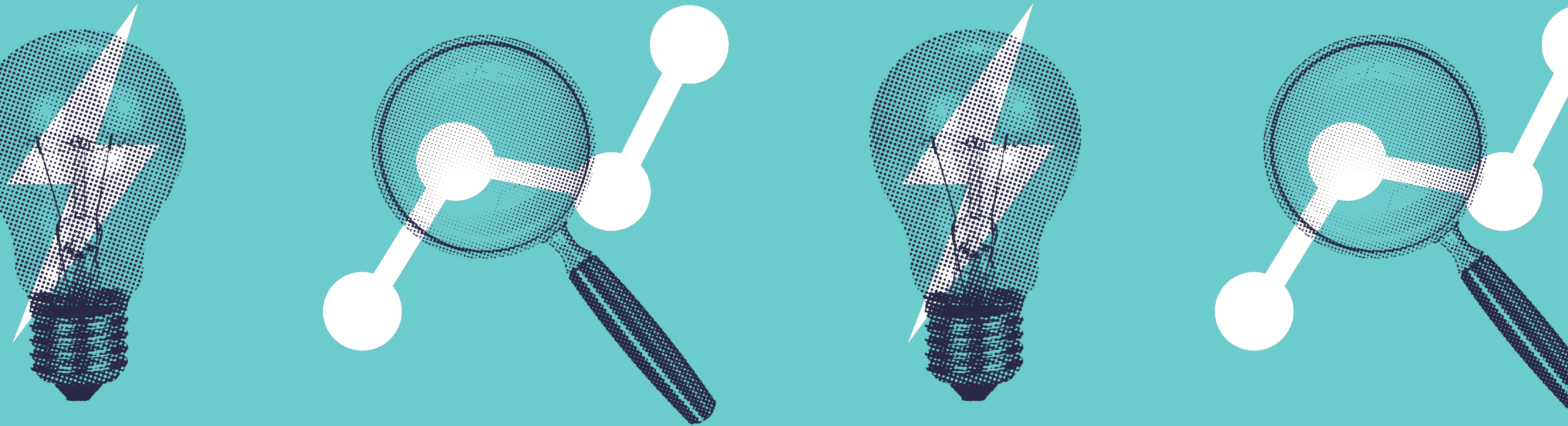
Measures to reduce the spread of misinformation on WhatsApp

WhatsApp has taken a number of steps to reduce misinformation, including making it easier for users to launch a quick search based on the message they receive, to encourage informed decisions about what user's are engaging with.

WhatsApp has also limited the speed at which forwarded messages can be shared: for messages that have been forwarded more than 5 times, users can only forward to one chat at a time. This is another example of **friction** in platform design to encourage different user behaviours.

Transparency

Centre for
Data Ethics
and Innovation



Transparency

- Transparency measures can increase our understanding of the policies and processes platforms put in place to address misinformation, and crucially the impact and effectiveness of these measures. **This can allow external scrutiny,** and is important in **increasing public trust and understanding** of the processes that impact user's activities online.

- In the UK, there are currently no legal requirements for transparency measures in content moderation and design choices. **Most platforms publish transparency reports on a voluntary basis**, which provide information on enforcement of community standards, usually including numbers of content and account removals and numbers of successful appeals against content decisions.

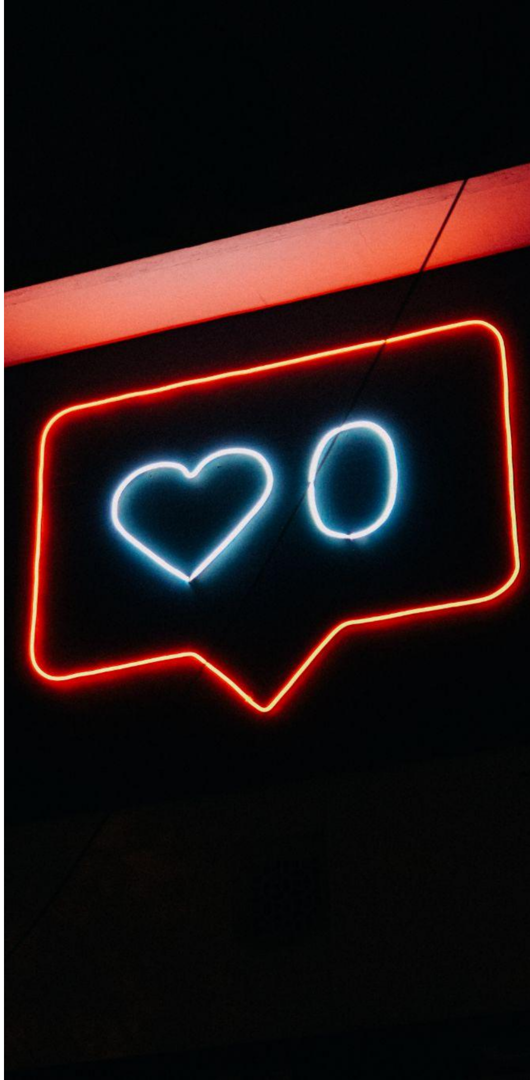
- While these can be helpful, **transparency reports often provide limited detail** across important areas including content policies, content moderation processes, the role of algorithms in moderation and design choices, and the impact of content decisions. Although many platforms have begun to provide more detailed information on their efforts to address misinformation, this is not always included in transparency reports, and can be piecemeal, varying significantly and presenting limited value to users and key stakeholders.

Content provenance verification

Manipulated content that claims to have come from a trusted news source can be particularly damaging and reduce trust in media organisations. Content provenance verification such as the work underway by the Coalition for Content Provenance and Authenticity (CP2A) could serve as a one way of ensuring publishers, platforms and social media companies know that content is coming from where it says it is, and that it has not been manipulated. CP2A sets open standards to develop end-to-end technical specifications on content provenance and authentication. The project aims ultimately to improve trust in the media through verified sources of information.

Transparency

- Beyond transparency reports, platforms have introduced new measures with the aim of increasing oversight and transparency.
    - **TikTok's Transparency and Accountability Center** aims to provide outside experts with an opportunity to see first hand how moderators apply guidelines to content flagged by algorithms, as well as being able to see source code. TikTok sees this as industry best practice that could and should be emulated by other platforms.
    - **Facebook** has introduced an **Oversight Board** to make final decisions on prominent moderation decisions. The Oversight Board recently upheld Facebook's decision to suspend former US president Donald Trump from Facebook and Instagram. The Board noted that **Facebook did not answer questions about the role of algorithms and design decisions in amplifying Trump's posts, making it difficult to assess whether measures short of suspension would have been sufficient** to reduce the risk of harm presented by Trump's posts.
- However, while increases in transparency are welcome, **participants questioned whether these are sufficiently meaningful** to allow necessary understanding and scrutiny of content decisions.



Transparency

- **Ofcom's role as the UK's online harms regulator** presents an opportunity to ensure greater transparency in the ways that platforms address misinformation.
- The new **Online Safety Bill establishes a duty of care on online companies to improve the safety of their users online,** overseen by Ofcom, with the largest platforms having a duty of care with respect to harmful but lawful content. This will include an obligation for platforms to **produce yearly transparency reports on how they are addressing harmful content**, with requirements to be set out by Ofcom in public notices.
- The Bill would also require Ofcom to establish an **advisory committee on disinformation and misinformation**, to advise on how regulated services should address disinformation and misinformation, how Ofcom should make use of transparency reports in this context, and how media literacy could be improved and supported.

CDEI's review of online targeting

In our *review of online targeting*, we recommended that the online harms regulator should have the power to require platforms to give independent researchers secure access to their data where it is needed for research of significant potential importance to public policy.

Under the Online Safety Bill, Ofcom must produce a report exploring access to information for research into online safety matters, and the extent to which greater access might be achieved. In preparing the report, Ofcom will consult the CDEI.

Transparency

Participants in our discussion highlighted a number of areas where the additional disclosure of information from platforms would be helpful:

- **The precision and recall of detection algorithms:** How many of the flagged items were flagged correctly, and how many were false positives (precision)? How do platforms measure the proportion of rule-breaking content that is actually flagged by algorithms, rather than going undetected (recall)? Platforms could provide a mix of data to help answer these questions, including broad precision and recall data alongside samples of data that are identified, misidentified or missed. Platforms could also open up their systems to trusted stakeholders to run their own queries.

- **How algorithms are trained and deployed:** What is the internal QA and QC process for algorithms before being deployed? How are they deployed initially, e.g. to particular groups, and why? What measures are taken to minimise the risk of bias in algorithms in content moderation and recommendation?

- **Metrics for measuring outcomes:** platforms have highlighted examples of how actions to address misinformation have reduced visibility and spread, but participants thought they could do more to explain this in a consistent and in-depth manner.

The Online Safety Data Initiative and access to online harms data

The differences in how online harms and related data are categorised and stored across platforms creates barriers to development of systems for identifying and removing harmful content online, particularly for technology that might be employed by smaller service providers. The Department for Digital, Culture, Media and Sport's Online Safety Data Initiative aims to facilitate better access to data relating to online harms, working with technology companies, service providers, civil society and academics, as well as the CDEI.

Transparency

- **Clarity on content defined as borderline rule-breaking:** YouTube and Facebook both take action to downrank borderline content, a category which is useful in the context of misinformation where content may not always be clearly rule-breaking, but still presents a risk of harm. However, the category of borderline content blurs the distinction between permissible and rule-breaking content, creates uncertainty for users, and risks arbitrary decisions. Information on relevant policies and processes may give greater confidence to users and other stakeholders.
- The harm arising from misinformation may be most apparent where users are consistently exposed to it, or particularly vulnerable. Understanding the intensity of exposure alongside the visibility and spread of misinformation across a platform could help assess where the some of the greatest risks of harm lie.
- There will inevitably be limitations to the above questions, and much will depend on the level of detail offered by platforms. Ultimately, increased transparency cannot entirely inform us about the effects of moderation, recommendation and wider design choices on users' beliefs and behaviours. As mentioned above, further research alongside transparency measures could improve understanding.

Collaborative forums for addressing harmful content online

Facebook, Microsoft, Twitter and YouTube came together in 2017 to form the Global Internet Forum to Counter Terrorism, focusing on technological solutions, research and knowledge-sharing to further develop their ability to address extremist content online.

This demonstrates commitment and willingness from platforms to work together to tackle shared problems, although as of yet there is no similar forum for misinformation.

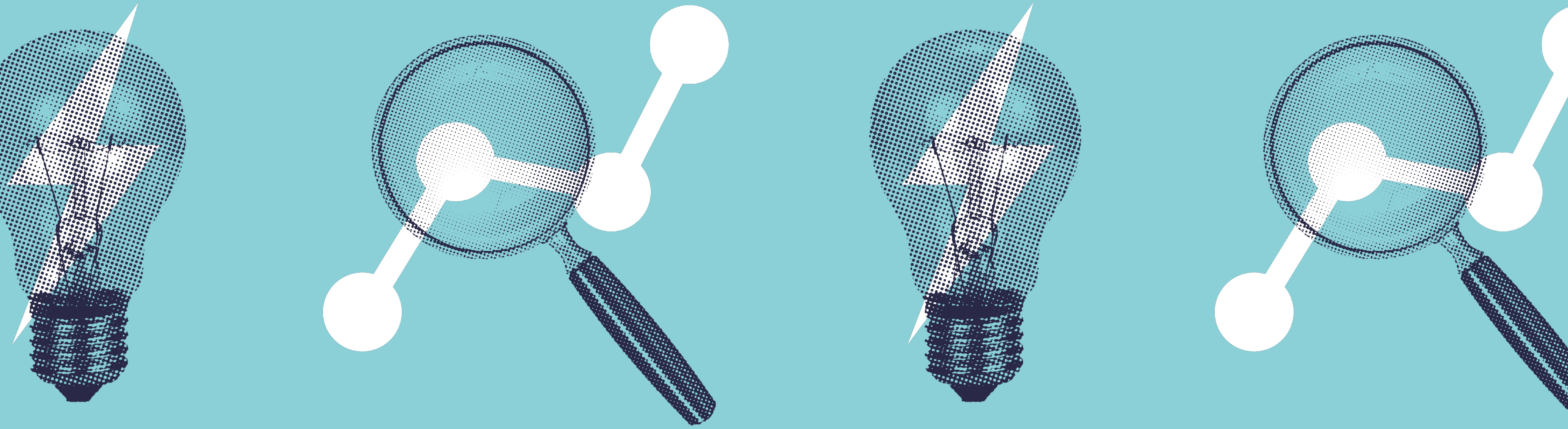
Transparency

- Guidance for increasing transparency will also need to consider a range of risks. Participants noted the risk of allowing malicious actors too much insight into the moderation process, which could allow them to avoid moderation, while platforms will also want to protect commercially sensitive information. Solutions have already been suggested for these problems, such as having arrangements for the sharing of particularly sensitive information with the regulator.

- Most platforms already produce annual transparency reports, but Ofcom has an opportunity to set a high but achievable minimum benchmark for transparency. Given the evolving nature of misinformation online and the urgent need to better understand its spread and effects, platforms should consider how they can report on misinformation policies and processes more regularly and consistently, even if less formal or detailed than the annual transparency report.

- While transparency alone is not a solution to the problem of misinformation online, it can help us understand where the greatest challenges lie, and where platforms may not be doing enough.

Conclusion



Conclusion

- Over the past year, misinformation has had a greater impact on people’s wellbeing, trust in democratic institutions and safety than ever before, increasing the spread of COVID-19, prompting the use of dangerous false cures, and hampering vaccine rollouts worldwide. In the face of these threats, and with human moderators in some cases unable to fulfil their usual duties, platforms chose to automate large parts of their content moderation process - if only for the duration of the pandemic.

- Our forum brought together experts to discuss the implications of this strategy. We sought to understand the different ways that algorithms can be deployed, their effectiveness in identifying and categorising harmful content, and what their role might be in the years to come, once the worst of the crisis has subsided. We also looked at what more should be done to improve transparency in content decisions and the moderation process.

- Few of our participants felt that fully automated content moderation could be a viable solution to misinformation in the near to medium term. Misinformation is diverse in nature, and its impact varies by context and audience. While algorithms can be effective in filtering and removing videos, images and text that have historically been flagged as harmful, they are less well suited to identifying novel forms of misinformation, struggling, for example, to distinguish between satire and material created with malicious intent.

Conclusion

- The platforms represented at our forum acknowledged these limitations, and have committed to returning to a pre-pandemic model whereby algorithms augment decisions taken by human moderators. However, even were machine-led content moderation to be viable, participants emphasised that there remains a normative question about what to do with the content that has been detected. Should it be removed entirely? Or should it remain on the platform but be labelled, downranked, and generally presented in a way that is more difficult to engage with?

- Participants discussed the merits and limitations of these different approaches, however noted that it is difficult to assess their efficacy without having more evidence available. Participants called on platforms to invest in more research to improve our collective understanding of methods like labelling and downranking. They also urged platforms to be more transparent about their content moderation policies, including how they use algorithms, and to issue more statistics, for example on content removal rates and successful appeals.

- Overall, participants in our discussion were pessimistic about our capacity to resolve the challenge of misinformation in the immediate future, with some believing that the prevailing business model of platforms was the greatest impediment to progress. However, our discussion showed that there are still many interventions that could be made today to constrain and mitigate the impact of harmful content, from supporting and promoting trustworthy media, to investing in valuable research on best practice, to experimenting with new technologies.