5RIGHTS FOUNDATION

# Risky by Design: Misinformation

May 2021

**Overview**

Misinformation is false or misleading information that can take many forms from memes to low-quality clickbait. Unlike disinformation, which is designed to deceive, misinformation is usually shared unintentionally. In practice, intention is rarely easy to determine and misinformation is often hard to identify.

Misinformation impacts large numbers of children with 55% of 12–15-year-olds reporting they have seen a false news story. This may have immediate consequences, for example research in June 2020 found one in five 16-24- year-olds thought there was no hard evidence coronavirus actually exists. Misinformation may also play a part in a slower but equally significant impact that influences children's relationship with the world. For example, 60% of children report they trust news less as a result of 'fake news'. This lack of trust permeates beyond news and information, to broader political processes and acts as a deterrent to civic engagement among young people.

Children have the right to access reliable information from a variety of sources. The impact of misinformation has been felt across society, but it has a particular effect on children who get more of their information online and may be unable to distinguish between what is true and what is false. When young people try to access credible information they encounter a digital world littered with misinformation.

Risky by design is a 5Rights Foundation project that shows how common design features of digital products and services create risks for young people.

**How do risky design features spread misinformation?**

Most digital services make money from advertising revenue and have a commercial incentive to keep users engaged. They aim to maximise the amount of time people spend using the service and in turn, the amount of data that can be gathered about them. This model sits behind the design decisions that ensure recommendation systems feed up content that is likely to provoke an emotional response in users and engagement in the form of likes, comments and views, regardless of whether that content is potentially harmful to users.

Dr Joan Donovan, Research Director or the Shorenstein Center at Harvard Kennedy School, recently told the US Senate "misinformation at scale is a feature of social media, not a bug," referring to the recommendation systems that serve up content based on a user's previous likes, shares and engagement, leading them down a content rabbit hole. When this repetition occurs across different services, narratives associated with conspiracy theories and misinformation appear more credible.[1]

[1] Statement of Joan Donovan, PhD. Research director at Harvard Kennedy school's Shorenstein Center on Media, Politics and Public Policy. Hearing on: "Algorithms and amplification: How social media platforms' design choices shape our discourse and our minds". Before the Senate Committee on the Judiciary Subcommittee on Privacy. Technology and the Law. April 27th, 2021. Available online here.

In a new case study, 5Rights highlights **eight** of the features that contribute to the spread of misinformation.

1. **Popularity metrics** likes, shares and views, inform the recommendation algorithms that digital services use to promote content to users. Misinformation may attract thousands or millions of likes, shares and views, particularly when it is provocative, humorous or even just absurd. Misinformation can seem more credible when it appears alongside visible popularity metrics or is shared by 'verified' accounts. Stemming the flow of misinformation is challenging when algorithms prioritise popularity metrics over the nature of the content itself, which can lead to misinformation being amplified and services profiting from its spread.

> The Centre for Countering Digital Hate (CCDH) reports; *"Social platforms chose not to alienate an anti-vaxx user base that CCDH estimate is worth up to $1 billion a year to them. Some platforms have even broken their own promises by still profiting directly from anti-vaxx content"*

2. **Recommendation systems** suggest content based on a user's previous engagement or the interests of similar users. When combined with popularity metrics, recommendation systems supercharge the spread of extreme content. This leads to a narrowing of subject matter and an increase in more sensational, often more egregious posts to view, videos to watch or groups to join. Algorithms use signals to inform what content users are recommended, such as who posted the content; when it was posted; whether it's a photo, video, or link; and the number of likes, shares and views it has amassed. The algorithms use these signals to predict how likely content is to be relevant and meaningful to you: for example, how likely you might be to like it or find that viewing it was worth your time."[2]

> A 2016 presentation from a major social media company revealed; *"64% of all extremist group joins are due to our recommendation tools"* and that most of the activity came from the platform's *"Groups You Should Join"* and *"Discover" algorithms: "Our recommendation systems grow the problem."*

3. **Autoplay** is designed to prolong time spent on the service. Services that use autoplay expose users to recommended video or audio content that plays without initiation from the user. Autoplay risks taking users further into recommendation rabbit holes and exposing them to recommended video or audio content that can become increasingly more extreme. Some services do not allow users to switch autoplay off.

> During the March to July 2020 UK lockdown as a result of Covid-19, young people told Ofcom[3] that they disengaged with the news but continued to receive

[2] Testimony of Monika Bickert, Vice President of Content Policy, Facebook. Hearing on: "Algorithms and amplification: How social media platforms' design choices shape our discourse and our minds". Before the Senate Committee on the Judiciary Subcommittee on Privacy. Technology and the Law. April 27th, 2021. Available online here.

[3] Ofcom Children's Media Lives: Life in Lockdown. August 2020. Revealing Reality. Methodology overview: "Interviews were conducted over six weeks from May to July 2020, so the children had been in lockdown for between 12 and 18

information via their social media networks.

On a service used by three quarters of 5-15 year olds, 70% of videos are viewed as a direct result of the recommendation algorithm.

4. **Trending lists** provide instant access to false information, particularly as popular hashtags are used to promote disinformation. Trending lists are easily manipulated by fake accounts and some companies exploit this by offering the creation of a "bot" account for as little as £150 to make a hashtag trend for a few hours.

An anti-vaccine video posted in April 2020 using #vaccine received 66,000 views on a livestreaming service popular with young people. The video is still viewable on the site today and is reportedly 'one of the first few that shows up' when you search for #vaccine, which itself has 42 million views.

5. **Fake accounts** include automated accounts or 'bots' and fake profiles created by users. Fake accounts can be created for malicious purposes, such as manipulating discussion online or spreading misinformation at scale. Bots that use AI to appear more human-like are difficult to detect and can evade content moderation. Despite policies to tackle coordinated inauthentic behaviour, fake accounts have been used to create pages where fake accounts can generate fake engagement. A loophole in inauthentic behaviour policy means that, using pages, malicious actors are able to exaggerate the credibility of information users see and ultimately influence the algorithms that recommend content to users.

During the United States' withdrawal from the Paris Agreement[4], 9.5% of the total number of accounts tweeting about climate change were considered likely to be bots, but these 'fake' users accounted for 25% of the total tweets about climate change on most days.

In a 2021 Guardian investigation, a former Silicon Valley employee describes a loophole which enables the creation of dummy pages to drive fake engagement in favour of political figures. Describing the scale of the problem, they explain in one case; "Over one six-week period from June to July 2018, [...] posts received likes from 59,100 users, more than 78% of which were not real people." Despite this, the service continues to focus on fake profiles, it "does not have teams that look for fake pages actively".

6. **Ineffective content labelling** undermines efforts to identify misinformation or provide relevant information to users. Content labels that warn of inaccurate content or redirect users to credible sources of information are often too subtle and

weeks and most were not attending school or leaving their homes to socialise. During these interviews, we saw the gradual 'easing up' of the lockdown restrictions, with slight differences in timings and guidance between England, Scotland and Wales." Available online here.

[4] According to Scientific American, "Thomas Marlow, a postdoctoral researcher at the New York University, Abu Dhabi, campus, and his co-authors" "measured the influence of bots on Twitter's climate conversation by analyzing 6.8 million tweets sent by 1.6 million users between May and June 2017."

therefore ineffective. Labels have also inadvertently led to disputed content receiving more attention. Visual warnings have been called for to overcome concerns about 'language and cultural barriers' that play a part in false information.

> MIT Technology Review reported that Zignal Labs saw the sharing of content that had been blocked increase "from about 5,500 shares every 15 minutes to about 10,000."

7. **Disappearing content** promotes disinhibited behaviour and 'consequence-free' content sharing. This type of content can only be viewed and shared during a certain time period and often 'disappears' before it can be fact-checked. Features that allow users to create disappearing content are popular among young people, but are also more difficult to report.

> 86% of 13-17 year olds use disappearing content to interact with their friends.

8. **Seamless sharing** occurs when online services provide ready-made, recommended list of contacts to share content with. Misinformation spreads fast in private messaging channels where sharing is particularly easy. The 'ready-made' nature of sharing, promotes rather than inhibits the spread of false information.

> Young people describe sharing news via screenshots or pasting a link into private messages, which may cause vital context to be lost.

Misinformation is often characterised as the work of 'bad actors' and many believe the solution is to enable users to identify false information. This case study looks beyond these narratives to consider the design features that increase the spread and reach of misinformation across services that are optimised to capture attention and extend engagement.

**About 5Rights Foundation**

5Rights develops new policy, creates innovative frameworks, develops technical standards, publishes research, challenges received narratives and ensures that children's rights and needs are recognised and prioritised in the digital world. While 5Rights works exclusively on behalf of and with children and young people under 18, our solutions and strategies are relevant to many other communities.

Our focus is on implementable change and our work is cited and used widely around the world. We work with governments, inter-governmental institutions, professional associations, academics, businesses, and children, so that digital products and services can impact positively on the lived experiences of young people.