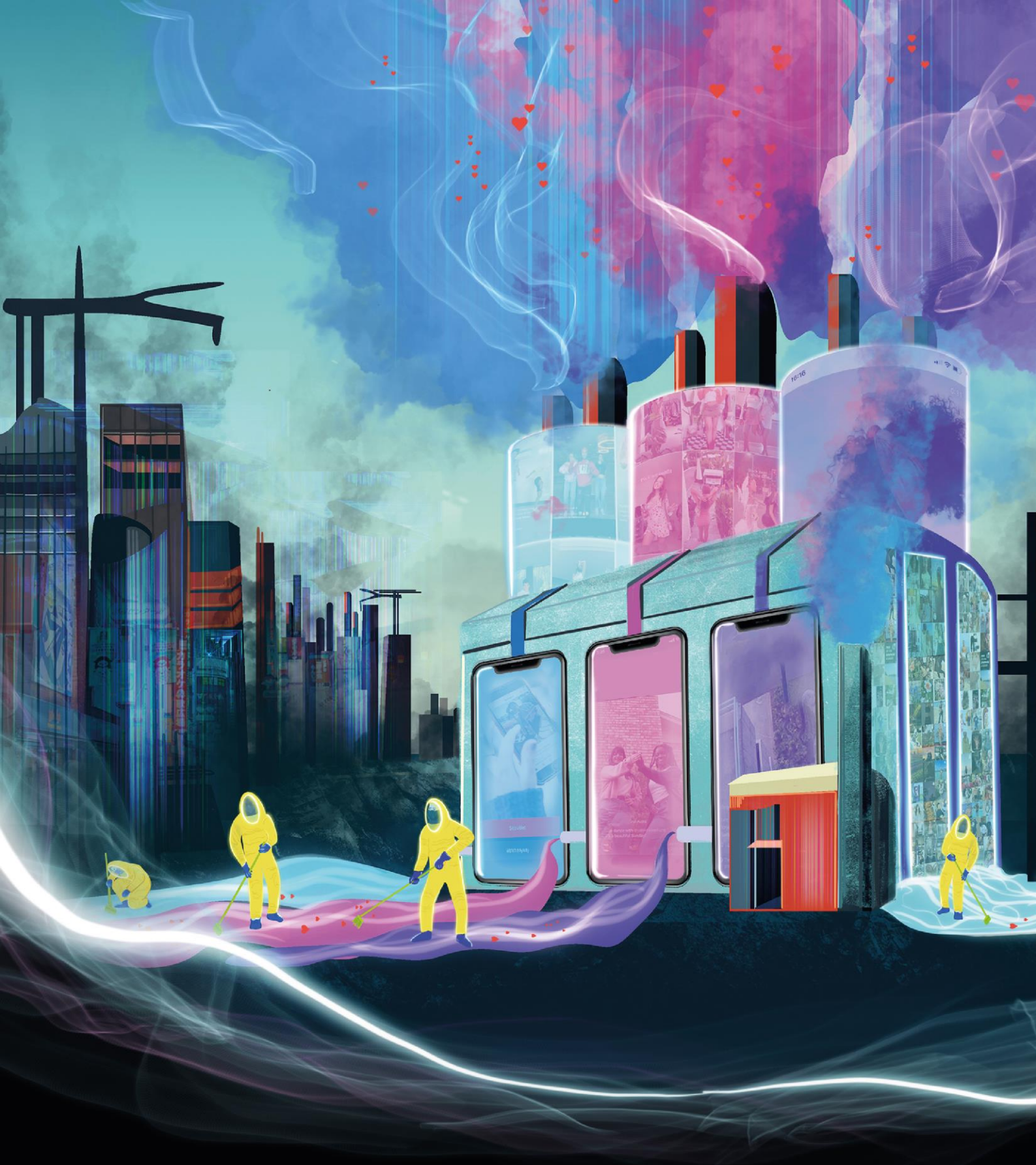


# Cleaning up in the attention economy

## Research into moderators' experiences

REVEALING REALITY

# About Revealing Reality

Revealing Reality is a multi-award-winning insight and innovation agency. We enjoy working on challenging projects with social purpose to inform policy, design and behaviour change. These include researching how digital services and platforms are shaping people's behaviour across relationships, media literacy, health, gambling, financial products and more.

We conduct detailed qualitative and quantitative research to build an in-depth understanding of digital behaviours and observe how people really experience technology. This has enabled us to become thought-leaders in online media habits and behaviours.

Over the last few years, we have conducted several projects specifically exploring digital design and online harm. These include understanding how digital products shape the lives and experiences of children for the 5Rights Foundation[1]; talking to young people about why they use porn for BBFC[2]; and exploring how adults are harmed online[3] and the risk factors that affect online harm to children[4] for Ofcom.

We have also researched families' attitudes to online age assurance for Ofcom and the Information Commissioner's Office[5]; worked with Internet Matters to develop its index on the impact of the digital world on children's wellbeing[6]; and been privileged over many years to conduct Ofcom's ambitious longitudinal study Children's Media Lives[7], which tracks children's behaviour, attitudes and experiences.

From time to time, we produce reports on projects we have funded ourselves, usually where we have observed something in the course of commissioned work that we think would benefit from further investigation. Recent examples include Not Just Flirting, a large quantitative and qualitative study of teenagers' experiences of nude-image sharing, which we did in partnership with the PSHE[8], and Through the Looking Glass, which explored how smartphones influence behaviour[9].

Revealing Reality has a strict ethics and safeguarding policy to ensure the safety and wellbeing of people taking part in our research. This policy is reviewed regularly to ensure it is in line with industry standards, including those of the Market Research Society and the Government Social Research Service.

Visit www.revealingreality.co.uk to find out more about our work or to get in touch.

---

1 How digital design puts children at risk, 5Rights

2 What do I do? How children use porn to explore intimacy, BBFC

3 How people are harmed online, Ofcom

4 Research into risk factors that may lead children to harm online, Ofcom

5 Families' attitudes towards age assurance, Ofcom and the ICO

6 Children's Wellbeing in a Digital World, Internet Matters

7 Children's Media Lives, Waves 1-8, Ofcom

8 Not just flirting, Revealing Reality

9 Through the Looking Glass, Revealing Reality

# Foreword
# 'Bad' content keeps 'coming back'

It was in the course of another project that we initially came across a social media moderator's perception that 'bad' content online keeps 'coming back' – that moderation often feels like a dystopian digital version of Whac-a-mole. The grimmest videos or images keep popping back up on another platform, or posted by a different user, or tagged with a different set of words. The job of the moderator was to clean up digital pollution after the fact, but it was a near-impossible task.

Moderators' voices have been largely unheard in descriptions of what happens online, particularly how harm occurs and the motivations and circumstances that drive people's online behaviour. Revealing Reality and others have done research in which we have sought to understand the experiences of users, and the motivations of digital designers. There have also been articles in the press about some moderators' experiences,[10,11,12] including those who were suing the platforms that had employed them for the effects of PTSD. But there has been no research putting moderators' experiences at the heart of an exploration of how and why the *system* operates the way it does in the context of what we already know from other research.

So we set out to do a small, self-funded research project in which we would find out if other moderators had the same experiences as the one above, and whether they had reached the same conclusions. Moderators not only see the content that users create, post and share, they also get first-hand experience of what platforms allow, promote and prevent. Their perspective is a valuable addition to developing an understanding of what is happening and why. We hoped exploring their experiences might generate insights that would be valuable to policy-makers, platforms and the public as they continue to debate what online harm looks like, who is affected and how and why, and what should be done about it.

One of the solutions often put forward to address the problem of online harm is greater use of artificial intelligence (AI). Most of the biggest social media and video-sharing platforms have publicly stated their commitment to increase the use of AI to attempt to reduce the potentially harmful content or experiences users have online. AI can be used to help detect markers that are likely to indicate content is of a particular kind, for example that it contravenes the platforms' guidelines. But its accuracy and applications are limited, and it is unlikely to provide a full safety net – certainly not in the near future.[13,14,15,16]

In the meantime, a huge amount of moderation is done by people. Social media and video-sharing platforms directly or indirectly employ staff to view and filter content that has already been posted by users. Their moderation influences what happens to content – who it is promoted to as well as whether it is removed. Moderators can also seek to influence the behaviour of users or shut down accounts that are in breach of platforms' policies.

But while there is a near infinite amount of content on platforms, there are inevitably only a finite number of moderators who are working to attempt to 'clean' it up.

Many moderators have signed agreements with their employers or former employers saying they will not divulge details of their jobs or experiences at work. So interviewing them for research is challenging, and it's essential we protect their anonymity. For these reasons, we have not only given the respondents pseudonyms, we have also changed some of their personal details so they cannot be identified, and chosen not to reveal their employer.

In this project we interviewed five moderators at length. While the sample cannot be said to be representative of all moderators, and it's not possible to extrapolate the specific findings, much of what we found relates to the ecosystem in which most social media platforms operate. This suggests likely broader implications.

---

[10] AI won't relieve the misery of Facebook's human moderators, The Verge, February 2019

[11] TikTok's content moderators watched beheadings, child sexual abuse, and other horrors. The aftermath is messy – and they say the company doesn't care, Business Insider, June 2022

[12] Facebook's content moderators are fighting back, Wired, June 2021

[13] AI should augment human intelligence not replace it, David De Cremer and Garry Kasparov, Harvard Business Review, March 2021

[14] AI won't relieve the misery of Facebook's human moderators, James Vincent, The Verge, February 2019

[15] Gillespie, T. (2020). Content moderation, AI, and the question of scale. Big Data & Society, 7(2).

[16] AI is not smart enough to solve Meta's content-policing problems, whistleblowers say, Business Insider, June 2022

# Executive summary

Most of us see the content in our social media feeds and assume it looks much the same as other people's. Even if we know intellectually that this isn't the case, we can't break free of algorithms and echo chambers – so it's hard to imagine accurately what we don't see ourselves. This difficulty seeing what others see adds to the challenges inherent in agreeing how to ensure people have positive experiences online, are protected from harm and are free to express themselves.

In contrast, platform moderators not only see the gamut of content that users create, post and share, they also get first-hand experience of what platforms allow, promote and prevent. They have a unique perspective.

**Moderation cannot eliminate online harm**

'Moderation' of content is often put forward as the solution to prevent online harm. Although many of the platforms tout their plans to increase and improve the use of AI to do this, others say that human moderators can't be bettered, certainly not any time soon[17].

However, this research highlights that not only does a reliance on human moderation in many cases 'transfer' harm onto moderators themselves, often it does not or cannot entirely or permanently remove content – certainly from the wider online space – even after moderators have deemed it inappropriate or, in many cases, distressing, disturbing or traumatising.

**Moderators' experiences provide insight into the workings of the attention economy**

While it must be stressed that this is a small sample, and that the people who were prepared to be interviewed may not be representative of all moderators, their experiences and perceptions nonetheless provide insight into what the job involves, and what it might tell us about the ways the attention economy – in which businesses buy, sell and compete for our attention – operates in practice. This includes exploration of the externalities of this economy.

We present the findings from this research, which is based on in-depth anonymised interviews with five former or current moderators at one video-sharing platform, in two parts.

**Moderators found their jobs hard and unpleasant**

One part (pages 8-14) describes the moderators' day-to-day experiences of doing their jobs. In common with descriptions elsewhere, this research has found moderators:

- found their roles unpleasant, stressful and sometimes traumatising;
- saw graphic and explicit content, including depictions of violence, murder, suicide, pornography, fetishes and self-harm to adults and children;
- did not feel they had access to adequate mental health support;
- felt their employer was less concerned about their well-being than about maintaining the flow of attention-seeking content.

**Moderators did not believe online harm would be eliminated while the attention economy operates as it does**

The second part (pages 15-19) reveals how these moderators not only felt let down by the platform that employed them, they felt a **bleak pessimism about the possibility for potentially harmful content, experiences and behaviour to be effectively reduced using moderation**.

This was because, as they saw it:

- The incentives within the attention economy for people to create, post and re-post content, including abhorrent content, over-rode attempts to remove such content.

  This was for two reasons:

  1. In their experience, content that garnered a great deal of attention, especially if it was shocking, would nearly always be re-filmed and/or re-posted, if not on the same platform, on a different one.

[17] Gillespie, T. (2020). Content moderation, AI, and the question of scale. Big Data & Society, 7(2).

2. At the same time, the supply of fresh content was ceaseless.

- Attempting to remove content after it had been posted rather than preventing it being posted in the first place maintained the incentives to create and post attention-seeking content and made it inevitable *some* content would harm *some* people. This harm can be seen as an externality of the attention economy.

- The platform did not want more content to be removed or restricted than was necessary, and the moderators sometimes believed these thresholds were too high to prevent harm.

*This report contains descriptions of content, experiences and allusions to topics that some people might find distressing. These include moderators' descriptions of graphic violent or sexual content they had seen in the course of their work, which they had found disturbing or upsetting.*

# What we did

## The aims of the research

Revealing Reality researchers have carried out numerous projects exploring children's and adults' experiences and behaviours on social media and video-sharing platforms, including exploration of what constitutes online harm and how it can come about.

Moderation is supposed to keep people safe from harm. But does it work? Despite many platforms putting in place measures to moderate content, we see across our work that hazardous – and sometimes illegal – content often remains on platforms long enough, and is shared widely enough, that it can reach a large number of people.

We sought to explore moderators' experiences and perspectives to gain insight into what their jobs involved, how the system operates, and how it could be improved.

Focusing on moderation at one major video-sharing platform, we set out to understand:

1. Moderators' **experiences** of their roles, including –
    - What does the role of content moderator involve?
    - What kinds of content do they see?
    - What impact does moderating content have on the moderators themselves?

2. Moderators' **perceptions of what drives users' behaviour**, including –
    - Why do they think people post shocking, disturbing or distressing content?
    - What role does moderation play in keeping users safe?
    - How effective is it?
    - How do they feel about their role as moderators?
    - Why does some content still make its way around social media when it appears to contravene platforms' policies?

We set out to interview people who were working or had recently worked as moderators to learn about what they were seeing and doing, how they were briefed, trained and supported by the platforms that employed them, and what was their view of moderation after some time in the job.

## Recruiting moderators to take part in the research

Recruiting moderators to interview was easier said than done. Many moderators – particularly those who were still employed in those roles – were worried or uncomfortable about speaking about their work or had signed agreements prohibiting them from sharing anything about their role, the content they saw, or the companies they worked for.

We trawled forums and LinkedIn to find individuals in moderation roles – approaching well over 100 online to see if they would be interested in taking part in the project. Many did not respond or told us they could not share details of their experiences with us.

As a result of these recruitment challenges, we narrowed our sample so that all respondents for this project were working or had recently been working in moderator roles at the same video-sharing platform. We learned that at this platform there were a variety of moderation roles. Round one moderators are the first human line of defence against harmful content, after an initial filtration of content using AI. Round two moderators reviewed content that had not been filtered out by AI or during round one. For this project, finding it impossible to speak to round one moderators, perhaps because they did not work in Europe, we focused on speaking to those working in round two.

### Interviews

We carried out interviews between March and July 2022 with five round two moderators who worked or had worked at the same video-sharing social media platform.

Interviews, lasting approximately one hour, took place remotely via video call.

At Revealing Reality, when we do qualitative research we not only ask research participants about the topic we are researching, but also gather information about them and their lives, so we can contextualise what they are telling us. This allows us to draw more meaningful insight from the research evidence.

As well as asking for this contextual information, during the interviews we asked the respondents about:

- Their experience of work – previous jobs and experience
- How and why they began working as a moderator
- The organisational culture at the platform
- How the platform's business model affects moderators and users
- Day-to-day experience of working as a moderator
- Their experience of the content they were moderating – what content was posted / reposted / viewed / liked / shared and why
- How the content affected them / their colleagues / platform users.

To maintain their anonymity, we have given the five research participants pseudonyms and we are not naming the platform where they worked.

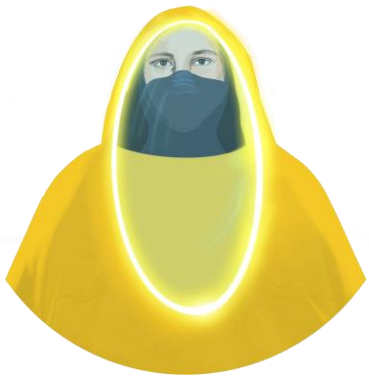
Certain details of their stories have also been changed so that they still reflect relevant contextual information but do not identify the respondents.

# Moderators' day-to-day experiences

To maintain their anonymity, we have given the five research participants pseudonyms and we are not naming the platform where they worked. Certain details of their stories have also been changed so that they still reflect relevant contextual information but do not identify the respondents.

---

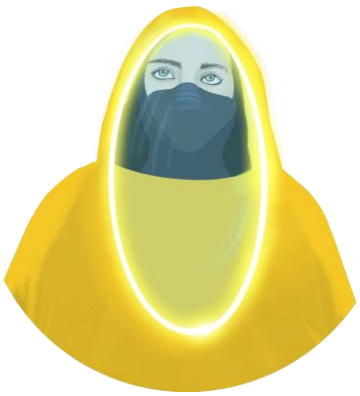
## Meet the moderators



**Nicolas,** who is in his 40s, moved to the UK from the Netherlands after he was offered the job at the video-sharing platform in late 2019, for which his salary was around £25,000. The ad for the job had mentioned 'content management' rather than 'moderation'. Initially, he had worked in the company's office, but during and after the Covid pandemic he worked from home.

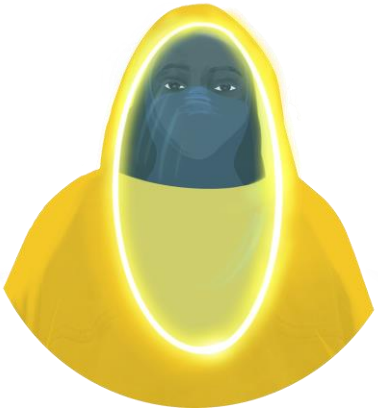
He had previously worked in the Netherlands in another content moderation role for a different platform and on a temporary contract writing online content for a travel agency in Austria.

Nicolas had taken several months' sick leave while employed by the video-sharing platform, which had included stays in hospital with mental ill-health. At the time we interviewed him, he was about to return to his job saying he needed the paycheque, but he was also looking for a role elsewhere.



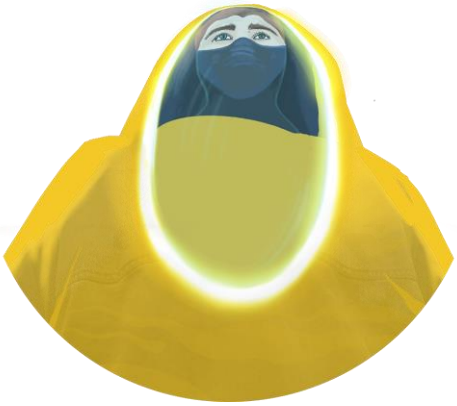
**Janine,** who is in her early 30s, is from France. After doing two years of a degree then deciding to take a break, she moved to the UK after a friend found her an opening as an au pair. She later decided not to go back to university, instead taking customer service roles, including one at an online travel agent, from which she was made redundant during the pandemic. Her salary when she joined the platform was around £24,000.

Janine worked for a year and a half as a moderator but resigned a few weeks before we interviewed her.

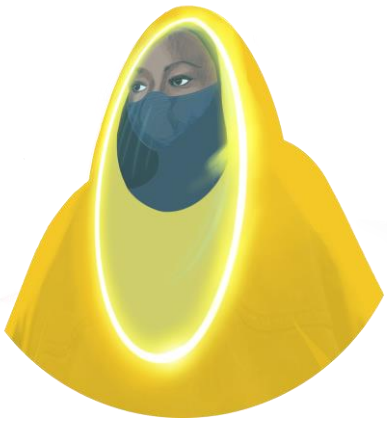


**Brigitte,** who was in her early 30s, used to work in the tourism industry. Originally from Italy, before the onset of Covid she had been working in a hotel in London. She was put on furlough during the pandemic and after a few months was contacted by a recruiter who said they'd seen her LinkedIn profile and began talking to her about a job working from home as a moderator at the video-sharing platform. Brigitte, who was "not doing anything anyway" was curious and thought it was worth a try.

She worked at the video-sharing platform for two years, starting on a salary of £24,000, which was increased by 6% after a year. She described the first year as "good" but left at the end of her second year because she was increasingly unhappy with the work itself and with what she described as the inflexibility of the shift patterns and annual leave.



**Christian,** who is in his early 30s, is Belgian and can also speak Italian. Christian studied in Italy before moving to the UK. He had been living and working in the UK for several years and became a moderator during the Covid pandemic after being let go from his previous job in marketing. He spent a year working at the platform, earning £24,000. He left because he felt the number of videos he was by then required to watch was impossibly high, he didn't have the support or the professional opportunities he had been hoping for, and he felt unable to fulfil his original "mission" to "combat" horrible content and make sure users did not see it. He returned to working in marketing.



**Caroline,** who is in her early 40s and Spanish, moved to Scotland after completing her Master's degree. She went on to study international relations in the UK before starting a job in logistics. During the Covid pandemic, Caroline lost her job and when searching for another one came across an advert for the video-sharing platform, which was looking for Spanish speakers. What most appealed about the job, which she was surprised to get "because I had no prior experience in moderation", was that she could work from home and it would be something different. At the time of the interview, she had been at the platform two years. Her salary had increased from £24,000 when she started to £25,000.

## Why did moderators take the job?

Many of the moderators we interviewed were drawn to their jobs by advertising that suggested they would be working for an exciting, quickly expanding video-sharing platform that provided the latest equipment in a flexible, work-from-home role. This included higher salaries than many of them were used to and a job which appeared to include status, responsibility, and the possibility of progression. Several of the moderators had been made redundant or put on furlough in their previous jobs during the Covid pandemic, which added to the appeal of the role at that time.

These round two moderators said they had understood before starting the job that they would not be required to moderate the 'worst' content because the AI systems and round one moderators would have already filtered it out.

If they felt overwhelmed by aspects of the job, moderators said they were told they could contact the employee assistance programme (EAP), which provided up to a total of 10 hours of counselling.

## What did the job involve?

The moderators we interviewed were 'round two' moderators. Before the video content reached them, it had already been filtered using AI and 'round one' human moderators.

As the round two moderators understood it:

1. **AI** automatically removed content that unambiguously contravened the platform's guidelines.
2. Human **round one moderators** reviewed still images from videos selected by the AI moderation process, and filtered out any further content they could see was illegal or not allowed by the platform.
   The round two moderators we interviewed believed round one moderation took place in countries such as China or India.
3. **Round two moderators** then watched videos in their entirety to tag them and flag anything problematic that had not been picked up by the AI or round one moderation.
   The round two moderators we interviewed were employed in the UK or Europe. One explained that round two moderators are often hired in part for their language skills, so they can make sense of the audio and cultural implications of videos alongside the visual elements.
4. A proportion of the tagging and flagging decisions by round two moderators are sent afterwards to a **"quality assurance" team**, which reviewed whether the moderators had made the correct decisions in the platform's opinion.

The moderators we spoke to said they worked:

- Entirely from home
- A 40-hour working week
- 8-hour shifts that started at 8am, 10am, 3pm or 11pm
- With a one-hour 'lunch' break plus additional breaks adding up to 30 minutes a day.

The round two moderators were responsible for:

- Assessing whether videos violated the platform's policies.
- Assessing whether profiles violated platform policies.
- Tagging videos by adding key words. These could range from "dog" to "teenager, girl, age 13-16, dancing" and were intended to help sort content so it could be 'promoted' or 'held back' on users' feeds as well as improving the content selection process of the AI moderation.
- Flagging videos which violated platform policies so that the video could be removed, passed along internally to another team, or remain on the platform but not be shown/promoted to users.
- Assessing whether situations were dangerous and alerting internal teams who could contact users involved and relevant authorities.

The round two moderators we interviewed had to:

- Moderate between 1,300 and 1,700 videos per day.
  Several of the moderators reported that this target had increased from closer to 1,000 videos per day when they had first started working for the platform.
- Check 'viral' videos – for example those that had more than 10,000 views – "to check that a video that is really popular is not popular for the wrong reason".
- Check particular "queues" of videos, as directed by their managers – for example queues of videos with a certain number of views, those that had been reported by users, or those in a particular language. Or work on certain projects, for example modifying the captions or notifications of videos that the platform might want to promote, or checking individual users were complying with platform guidelines.
- Consistently tag and flag videos correctly according to the platform's policies (see box, below).
- Justify or explain 'bad cases', which the quality control team concluded had been incorrectly tagged. One moderator told us too many bad cases reduced the amount of bonus they were paid.

The quality team would check approximately 10% of a round two moderator's video-tagging and flagging for accuracy.

**What happens to content that has been tagged or flagged?**

Content that had been flagged or given certain tags was not necessarily removed from the platform. Often the tags determined the degree to which content was or was not promoted, and to which users, though the moderators didn't control what happened to content after they had tagged it.

For example, content tagged with words such as "dog" or "funny" can be directed into the feeds of users deemed likely to engage with that kind of content based on their profile and their previous behaviour on the platform. Moderators said there were several hundred tags they could apply.

*"If you see the belly of the girl, you have to tag it. If you see the tights, you know, the legs, you have to tag the legs. If the dance is sexy, we have to tag 'sexy dance'. And there's a[nother] different level of sexy dance.*

*"Is that a dog that passed around the corner and you can see his genitals? We have to tag that, we have to make the difference [between] if it is intentionally in focus or unintentionally in focus."*
**Nicolas**

*"If people are filming something random, like filming someone walking down the street, we wouldn't push that. It has to be interactive, entertaining."*
**Caroline**

## What were the moderators' day-to-day experiences?

In contrast to the benefits of the job the moderators had been told about when they were recruited, in practice all five of the moderators we interviewed reported finding the job difficult, exhausting, upsetting and sometimes traumatic. There were several common themes and experiences.

### They found the work repetitive and dehumanising

Their hopes that the job would be fulfilling, as they worked to protect others, were quickly tempered. They felt "robotic", doing a high volume of work that was often monotonous but was expected to be done at speed.

*"At some point I was like, oh my god, I feel robotic… When you don't have to really listen to the content, when it's just some music that is, you know, like trendy music or something like that, you just moderate what you see and then you can launch another content at the same time and so on."*
**Christian**

### They found the volume of content they were expected to moderate unmanageable

The moderators we interviewed, who had all worked for the platform at a similar time, said they were expected to watch hundreds of videos a day. For some, this daily target had increased from around 1,000

videos when they first started, to approximately 1,700 videos months later. Though breaks were encouraged in theory, many of the moderators said the volume of content they were expected to get through each day was unmanageable if they also took breaks.

Each had developed tactics for getting through these high workloads. These included:

- Speeding videos up
- Watching multiple videos at once
- Choosing which videos to watch with or without sound

*"I watch around 30 and 40 hours of videos every day. Because then up to four videos at the same time and four times the speed. And you need to check the video and the audio."*
**Nicolas**

*"Certain kinds of video we know we can speed up. Otherwise it [meeting the target] wouldn't be possible. For example, cooking videos. You know, it's very unlikely that they're going to swear or [something], so that helps."*
**Caroline**

### They sometimes made mistakes

Because moderators felt they had to use these various techniques to get through their workloads, they inevitably missed things.

*"I passed the video times four. Like even the captions and everything was in Russian… And then my manager told me, maybe you should look at it another time. But without the speed and so on."*
**Christian**

Moderators were assessed on the number of "bad cases" in which they had tagged or flagged content differently from a co-worker. When this happened the moderators had to justify their actions, adding to the sense of pressure. The number of bad cases could also affect their bonus or progression.

Keeping abreast of the changing criteria by which content should be moderated was a challenge, especially given the pressure on them to later 'defend' moderating decisions that were at odds with colleagues'.

*"There are changes all the time, which is really hard to face sometimes. There are exceptions of exceptions sometimes. You know, it's like learning a language."*
**Christian**

### The moderators saw graphic and distressing content

Despite initial reassurances that they wouldn't be moderating the most explicit content, all the round two moderators we interviewed told the researchers they did frequently witness content they described as "extreme", "disturbing" or "shocking".

They were tasked with removing or flagging content that contravened the platform guidelines but were expecting most of this content to be relatively benign – and said this was the case for most of the content they viewed. Nonetheless, between them, the five respondents had seen videos showing murder, suicide, accidental death, porn, kidnapping, animal cruelty, extreme violence, and fetishism involving children.

*"I saw people hanging themselves. I saw a girl shooting her head off, the brain coming out, and her hands grabbing the couch when she had no head."*
**Nicolas**

*"There was one video, which was the one that most scarred me. It was two kids. They were playing with a nail gun and they were recording and the girl fires the gun, and you can only see the guy screaming and then the kid falling down. You don't see the nail getting shot like going through, but you can hear it. She [the other kid] does the same."*
**Janine**

*"There was a guy on [another social media platform] that committed suicide on 'live' [streaming]. So it was not a recorded video. And I know there was some article about it and when it came out obviously some people, some user of [the video-sharing platform] re-uploaded the video on [the platform]. So we had to see it again even if we knew it already happened."*
**Brigitte**

*"I saw the body, like some body parts that were floating sort of and I was like, okay, wow. So that was literally someone, you know, in a bathtub full of blood."*
**Christian**

### Some moderators found seeing graphic violent, shocking or sexual content traumatising

All five moderators reported being disturbed by some of the content they had seen. Two of them experienced longer-term anxiety and distress as a result.

*"These are images that you can't really forget, you know? They stuck."*
**Nicolas**

*"The animals and the people dying, those are the ones that for me got in my head for the longest, and I have to figure it out. It's with me until this day."*
**Janine**

*"I was petrified … That was horrible. I felt really, really bad seeing that"*
**Christian**

The consequences for moderators seeing disturbing pieces of content ranged from added stress during work (such as worrying the video may appear again), to needing additional support (such as therapy), to hospitalisation for mental illness. All of the moderators interviewed saw content which they found disturbing, affecting their experience of their jobs as well as their general mental health.

## What support were moderators given at work?

### Moderators did not feel able to take sufficient breaks when upset by content

All of the moderators we spoke to had seen graphic content that they found disturbing, but felt they weren't able to take sufficient breaks or measures to recover.

*"They tell us 'You can take a break, no worries, you can take five minutes' but at the end they manage our break. So we have only 10 minutes – three times 10 minutes – so if I already took my breaks and then I have a bad video, I want to take five minutes after, maybe the team leader will come to me and tell me, 'What happened there?' and then I have to explain myself, which is exhausting for me."*
**Brigitte**

*"We are the ones dealing with all the nasty stuff and it is tough at times… Even as you do [take breaks], you're mentally tired watching a thousand videos."*
**Caroline**

Three of the moderators also talked about the inflexibility of the working patterns – shifts at anti-social times and not being able to book annual leave when they needed it. Scheduling was at the discretion of the team leader, and more than one of the moderators we spoke to complained that their working pattern left them feeling isolated, alone and unable to take time off to rest and recharge when they felt they needed to.

### Therapeutic support was limited and difficult to fit into shifting schedules

Two of the moderators sought support to deal with the effects of the content they had seen or the experiences they had as moderators. A programme was in place – the employee assistance programme – through which the employees could access a limited number of therapy sessions.

*"I was in therapy for three weeks, and it was therapy like given by the people [external company contracted by the platform] …but I had to give it up because I couldn't do the therapy and work at the same time because they would collide with each other. Because I had shifts that started at 8am and finished at 5, they were working from 9 to 5."*
**Janine**

"*We have something called EAP [employee assistance programme]. Yeah, some programme when you can see a counsellor for 10 hours… online, Zoom meeting… one hour per week."*
**Nicolas**

The moderators who used it felt this limited support was not sufficient. Nicolas was on sick leave for five months to deal with mental health issues resulting from his work. He went to the hospital on three separate

occasions for mental health reasons during this time and said he received no further support from his employer.

*"And you know, it's five months. I'm out of work. And I've heard nothing about them. I just had one email from H.R. two weeks ago to tell me that I wasn't being paid any more. You know, not even a word from my team leader, who are supposed to take care of you. He doesn't even know if I have a brother, where I lived, what I did before…"*
**Nicolas**

*"I even had to take two months off work, because my brain would just be not switching on anything. I woke up crying on the days that I had to work. I didn't want to work. I, literally my head was a huge mess."*
**Janine**

### Three of the moderators left their roles within two years

Christian, Janine and Brigitte had all left their moderator jobs after less than two years.

Janine felt the platform had not provided adequate support.

"*So for them, it's just easy to train new people all the time, than keeping, retaining the ones they have… I thought, that's not worth it. You can't do anything, to be honest. They [the moderators] are always going to have that content, they're always going to see something."*
**Janine**

Nicolas had returned to his role after several months' sick leave, but said he was looking for jobs elsewhere. Many of his colleagues had already left, he said.

"Up to eight colleagues quit for mental health reasons – [names them]. They are gone because they were [driven] mad."
**Nicolas**

Caroline, who of the moderators we interviewed reported the least stress or distress as a result of her work, nonetheless said the job was "really tiring" and she felt she had "done enough". She had been unsuccessfully applying for other positions within the company and said if she didn't manage to get a different internal role within six months, she would look for a job outside the company.

# Moderation vs behaviour

Social media platforms differ in their policies, community guidelines and ethos, but have in common that they largely rely on moderation to help encourage the behaviour they 'want' and enforce the rules they have in place.

On the majority of platforms, moderation of content takes place *after* it has been posted on the platform, though moderation may also remove or suspend users who are seen to have repeatedly or deliberately breached rules or guidelines.

Sometimes, as is the case at the video-sharing platform in this research, some moderated content is still available to view by certain users but not others, for example those who are registered as children.

However, the moderators interviewed during this research were dubious about the extent to which their moderation could be effective. They felt that through a combination of:

- Inadequate policies;
- Inadequate or inconsistent application of policies;
- Insufficient numbers of moderators;
- Extremely high volume of content;

- Incentives for users to create or recreate ever more content,

moderation was a drop in the ocean compared with the tidal of wave of content. Their job – to 'clean up' the content on the platform once it was already online – was not only unpleasant and at times traumatic, it was, in fact, impossible.

This section outlines the conclusions the moderators had reached based on their experiences of moderation, the content they had seen and the ways the video-sharing platform operated.

## Removing content is only a temporary fix

### Extreme content that had been removed quickly reappeared

All the moderators saw content that had been taken down for violating guidelines quickly reappear on the platform, or resurface on a different platform.

*"I've seen a few contents that were a bit disturbing, would say murder-related. I've seen that it was some content that was taken from one platform and put on [social media platform A] out of nowhere. And someone filmed that on [social media platform B] …so it doesn't take long."*
**Christian**

No matter how quickly AI or the moderators act to remove content, moderators felt it was almost impossible to stop content being seen, copied and reposted:

*"So they [another platform] managed to ban the video, but not quick enough. You know how fast it is on social media. So, some people were able to actually share the video [on this platform]. My colleague, sadly, she came across that video and she actually took a few days off. She said, I can't take it. Because it was shared so many times, some other colleagues watched this video [too]."*
**Caroline**

*"The whole day I was dreading that the video would come up again. I thought, please, let no one put that video [up] again. And it's something that I can still tell you, like if I could film my head, I could literally play the whole video, it was that bad."*
**Janine**

## The design of the platform fuels the creation of more extreme content

### The moderators often recognised viral 'trends' in the content they saw

Moderators recalled a lot of the content being 'trend driven', observing similar patterns and themes across the content that was posted on the platform as users sought to emulate content that was proving popular.

*"Once someone starts a trend, then you see the same trend all day long."*
**Caroline**

Often the trends were benign – "boring" to have to watch, according to the moderators, but not always. Brigitte described a trend around a fetishism involving children as one of the worst videos she had seen:

*"It became kind of a trend, like we had many videos of it. It's with children… this is what confused me a lot because it's about… children taking videos of them peeing on themselves or shitting themselves but with their clothes [on] and it would be focused on their genitals, like, peeing."*
**Brigitte**

### Extreme content is rewarded with attention

Asked why they thought people were posting extreme content – graphic or shocking videos, often featuring violence or sexual activity – the moderators said they thought users were seeking attention or fame. This

drove their desire for engagement – posting content to attract more followers, likes and shares, and in the hope their post might go 'viral'.

*"That person is going to get attention, even if it's bad attention, or even if it's, like, 'compassion' attention. They still want to get it. I think that's what it is. Yeah, people have a problem nowadays and I think that problem is called attention."*
**Janine**

*"People just want to get viral and they want to get views. So it's disgusting… People want to be noticed. So they're willing to do anything just to have views, just so that people can share their content."*
**Caroline**

While moderating, Christian saw a graphic video of a murder onto which a user – not the original creator of the video – had added the username he used on another platform.

*"It's just probably to be seen, you know, under the spotlight. Like saying, 'Oh I posted it and this is my content because there was a banner with the guy's [social media] name and I was like, 'You really want people to talk to you on [name of another platform], like, because you posted a murder?!'"*
**Christian**

Caroline, too, said her conclusion was that "sadly, people are quite attracted to violence".

*"I think violence triggers something in your mind. There's some kind of weird fascination, seeing someone dying on a video. I don't know what it triggers into them, but when it comes to sharing, I think they just want to make a lot of views because they know that people are going to talk about it."*
**Caroline**

## Creators of content that gets a lot of attention are financially rewarded

Users' desire to post or share content that gets a lot of attention is enhanced by the prospect of financial reward, the moderators said.

The platform where these moderators worked offered a variety of ways users could make money if they posted content that lots of other users engaged with. These included:

- direct payments from the platform to users who have more than a certain number of followers and who post content that gets above a certain threshold of views
- virtual gifts from other users that can be converted into real money via the platform
- tips
- promotion of the user's own products or services
- paid promotion of other people's products or services
- affiliate marketing
- brand ambassadorships
- sponsorship

As these mechanisms were introduced or extended, three of the moderators said they noticed an increase in sexualised or shocking content, and content that emulated other posts that had garnered attention online.

*"That's when I noticed that there was a lot more sexualised content, because obviously, the more videos you make, the more money you earn… As soon as you show your boobs obviously you're going to get a lot of views, a lot of comments and more money.*"
**Caroline**

*"[The platform] became very weird. I saw a difference in the content, that it was more and more [designed] to be followed. Like at the end they always say 'Like my video'… after so many views and followers, you get paid for your video. So it became huge. At that moment, everybody wanted to be paid for their views. It's time to be viral at that moment."*
**Brigitte**

The moderators felt that despite having a moderation system to remove extreme content, the incentives – whether directly financial or simply to get attention – could motivate its creation, sharing and re-sharing. No matter how much they removed, there was always more being created or re-posted.

### Users are primed to record and share extreme content whenever it may surface

For these reasons, the moderators felt some people would film, create, share or re-post any content that might get attention, almost unthinkingly.

Even while watching livestreams, users can be ready to record, download, and distribute any extreme content that may crop up during the stream. In response to asking how content is posted if the user is dead (or, for example, commits suicide while filming), Nicolas explained:

*"Yeah, it's you know, some people, they kill themselves live on some platforms. So yes, and people copy the video and they put it on all the platforms so you can find it on [video-sharing platform] as well somewhere."*
**Nicolas**

Janine, too, described that even after the whole video of the two children getting injured with a nail gun had been removed, she kept seeing stills from the video that were "cropped and then posted" on the platform again.

Caroline had seen a video of someone being blown up after stepping on a mine. Asked what could be the motivation for the person who made the video to film and then post it, Caroline said she had asked herself the same question.

*"Was that accidental? I don't know if they knew. That's clearly a horrible, horrible thing to do… Just to get some views. You know, some people like to see that kind of conflict."*
**Caroline**

Christian described a video he'd seen of someone being murdered.

*"That was from a security camera… I think it's out of stupidity. Because basically we see so many movies, so many things that now I do not think that this person was intending to harm anyone. I think that it was just 'Oh, well, I see that happening. Um, I called the emergency [services], but I kind of record it just for my own record, and then sent it to other people.' However, to be put on the platform, I have absolutely no real idea of what was the intention. Apart from the fact that this person might be completely detached from reality."*
**Christian**

## Some moderators felt the threshold for removing content or users was too high

Several of the moderators stated their opinion that the threshold for removing certain content or users was too high, or that policies were applied inconsistently, in some cases meaning content was tagged differently. For example, Janine described that two women could be wearing the same outfit in their video, but if one of the women had bigger breasts, her video would need to be tagged to reflect that.

Janine also recounted coming across a video which showed a kidnapping taking place. She wanted to alert the authorities and take the video down, but both requests were denied on the grounds that she would not be able to prove the video was real rather than staged. She also saw a video where an animal was being harmed, but similarly was unable to prove that the animal was dead by the end of the video and so was not approved to take the video down.

### Moderators needed at least four videos to 'prove' a child was under 13

Brigitte complained that the level of evidence required to remove an account of a child who was under 13 was too high. She recounted that she used to go against platform policy and remove accounts where she felt satisfied the child was under the allowed age.

*"For the kids, it's not enough control, in my opinion… Sometimes people report 'I think he's less than 13' and so we will check the account. But we will need at least four videos in order to say I'm sure he's less than 13 and can delete the account. So sometimes I have three videos and it's a kid of eight years old just taking selfies. But I myself I will delete it because I see it's a kid. But if I follow the policies and the direction I shouldn't have deleted it."*
**Brigitte**

Brigitte also recalled that the platform used to moderate content posted by teenagers so that it would not be recommended to other users but this was no longer the policy.

### Moderators were told not to 'over-kill' the content

Janine thought the platform was purposefully allowing questionable content to stay online because it would gain attention, and in turn encourage others to create and post similar content.

*"Do not 'over-kill' it. That was the name they gave it. They didn't want to over-kill content. So if you take [down] too much, then you're going to over-regulate… I think it was for them to have, like, a lot of content on the platform."*
**Janine**

*"It's a bit of a stupid job, you know? But that, you know, that's how it works. If you run a company like [a video-sharing platform], you need to give the content that people want to see. And that's what is disturbing because there are really disturbing people all over the place."*
**Nicolas**

Janine described what she felt was a balancing act by the platform, where the tags the moderators have to apply serve to remove some material, while promoting other content.

*"I think they want to keep a balance, you know, like not having too much of one, but not having too much of the other either. So they try to maintain a line. I think it's [about] having as much content as possible… having endless content."*
**Janine**

# Afterword
# From attention to pollution

The five moderators we interviewed for this research project echoed what we'd heard from the very first one we spoke to. Based on their experiences at work, it was practically impossible to keep all the content on this social media platform 'clean'.

The moderators had their own opinions about why this was. Some of it was simply the sheer volume of content – there are only so many moderators, and the ones we spoke to felt it was impossible to keep up with the amount of content they had to check, and to do so without missing things or making mistakes.

Two of the moderators also explicitly questioned whether the platform really wanted all the content that people might find distressing to be removed. One had reached the conclusion that the moderators were not wanted to "over-kill" the content.

Whether that had been communicated explicitly by the platform or not, and exactly what it meant if so, it was clear that allowing platform users to post first, and moderators to check content after it is online, means any piece of content might be viewed by users before it is moderated.

Then there's the question of whether the moderators' own negative experiences are considered a price worth paying for this method of keeping parts of the internet 'safe'.

Because the sample for this research project is small, and the participants were self-selecting, we need to be alive to the fact that their experiences may not represent the experiences of all moderators, and certainly not all moderators at all platforms.

Nonetheless, based on this research and our work on a range of projects[18] exploring people's behaviours and experiences online, including previous interviews with people who design the platforms' functionality, it's possible to use the research evidence to draw inferences about the wider system. This system – the 'attention economy', in which businesses buy, sell and compete for our attention – is not new, but this research provides additional evidence that:

### 1. The motivations and incentives built into the attention economy shape people's behaviour

In the attention economy, many digital businesses' financial success relies on advertising revenue. Their profit is driven by maximising that revenue, which in turn relies on capturing – and then essentially selling – as much of people's attention as possible. In practice, this means these digital businesses seek to maximise the number of people who use their product, the time they spend using it and the depth of their engagement with it.

Inevitably, this business model drives the behaviour of those who work for these companies, determining priorities and influencing decisions. In previous research for 5Rights[19], we demonstrated how this shapes the way the features and functions of social media platforms are designed.

The products are engineered to maximise reach, time and interaction, with a ready-made audience on which to test features and tweak according to what delivers the greatest engagement.

Each individual platform is not only seeking to maximise the attention it gets on its own platform, it is also competing with other platforms who might otherwise capture that attention instead.

The designers themselves may not even be aware of the wider consequences of their design decisions – they are simply following instructions to reduce friction, or increase dwell time, or match a feature that a competitor has introduced.

These features and functions inevitably shape users' behaviour – incentivising people to create, post and re-share content that might capture other users' attention. This might be entertaining, informative or inspiring content, but equally it might be shocking, sexual, violent. The users are – perhaps unknowingly – supporting the platform's business model.

In practice, the attention economy has evolved based on what people are interested in and will pay attention to, which includes a mix of benign and hazardous content, rather than being actively designed to prompt or promote behaviour that might be less likely to cause harm.

### 2. The attention economy has externalities – indirect but large harmful costs borne by people who aren't in control of that system

In this research we can see that the negative experiences of the people whose job it is to moderate the content are one externality of this economy. Even if we assume that not all moderators feel as unhappy about their experience of moderating as the sample we have interviewed, we understand that at least some moderators are experiencing distress and sometimes mental ill-health.

But it's not only those who are trying to do the 'clean up' operation who are harmed. Because moderators on most social media platforms are required to check the content only after it is already live, any number of the

[18] E.g. Not Just Flirting Revealing Reality self-funded research; Pathways: How digital design puts children at risk, 5Rights; How People Are Harmed Online, Ofcom; Young People, Pornography and Age Verification, BBFC; Live Streaming: A new economy of connection, Revealing Reality self-funded research

[19] Pathways: How digital design puts children at risk, 5Rights

tens, hundreds or thousands of people who view the content before it is moderated may have negative experiences as well. This is also an externality of the system.

Then there is the question of what effect the attention economy – in particular the behaviours it incentivises – has on norms and culture more widely. What are the consequences for society if some people's first instinct when they see something horrific is to reach for their phones to film it and share it? The compulsion to film, share and reshare graphic acts of violence or horrific accidents to get views, likes, shares or followers, to unthinkingly or uncaringly 'pollute' the environment with content which seems often to override previously held collective notions of decency and sensitivity, is arguably an externality too.

It's easy to read the evidence in this report and conclude that there is no alternative, that the way the attention economy operates has both incentives and externalities built in. But, as history has shown, economic externalities can be tackled, once they are widely recognised and understood.

*If you would like to talk to us about the findings in this research project, or any of our other work, please don't hesitate to get in touch with our managing director Damon De Ionno damon.deionno@revealingreality.co.uk*