Algorithms Patrolling Content: Where's the Harm?

An empirical examination of Facebook shadow bans and their impact on users

MONICA HORTEN

22 February 2021

Working Paper

Abstract

At the heart of this paper is an examination of the colloquial concept of a 'shadow ban'. The paper reveals ways in which algorithms on the Facebook platform have the effect of suppressing content distribution without specifically targeting it for removal, and examines the consequential stifling of users' speech. A shadow ban is a term that refers to specific scenario where users' content is hidden or deprioritised without informing them. The paper reveals how the Facebook shadow ban works by blocking dissemination in News Feed. This is Facebook's recommender system that curates content for users, and is also the name of the algorithm that encodes the process. The decision-making criteria are based on 'behaviour', a term that relates to activity of the page that is identifiable through patterns in the data. It's a technique that is rooted in computer security, and raises questions about the balance between security and freedom of expression.

The paper is situated in the field of research that addresses the responsibility and accountability of the large online platforms with regard to content moderation. It works through the lens of the user to examine the impact of the Facebook shadow ban. Users, whether they are acting as speakers or as recipients of information, have positive rights that must be protected and they should not be treated as passive victims. The user experience was studied over the period of a year from November 2019 to November 2020 across 20 Facebook Pages from the UK. Data provided to the Pages via Facebook Insights was analysed in order to produce a comparative metric, and it was considered how the shadow ban could be assessed under human rights standards.

The paper concludes with a recommendation for quality controls on Facebook's internal processes, potentially with a form of triage to identify genuine, lawful content that has been caught up in the security net. Overall, an improved understanding should be developed around the automated processes and algorithms that are used in content moderation. This is a vital step to safeguarding the online platforms as a forum for public discourse.

About the Author

Dr Monica Horten is an independent scholar in the field of content online policy. She has served as a Council of Europe expert and Visiting Fellow London School of Economics. She is the author of three books in the field of content online policy: 'The Closing of the Net' (Polity2016); A Copyright Masquerade (Zed 2013) and The Copyright Enforcement Enigma (Palgrave 2012). She is a member of the EURALO Individuals Association - the ICANN At-Large (individual Internet user community) for the European region. Her website and blog is www.Iptegrity.com, and she tweets as @Iptegrity.

Contents

Introduction

This paper explores the case of UK-based Facebook Pages that unexpectedly noticed strange effects, in particular that they were significantly losing reach and their posts had become invisible, an effect that they described as 'ghosted'. From their perspective, they felt they had been put under some kind of mystery 'ban' but they had no idea what they had done to deserve it.

The paper is situated in the field of field of research that addresses the responsibility and accountability of the large online platforms with regard to content moderation. This is a relatively new field that has grown rapidly, in line with political concerns about the dominance of the platforms and their gatekeeping role for news, information and entertainment. A key issue is the increasing use of automation, artificial intelligence, and algorithmic decision-making along with concerns around freedom of expression and transparency. It addresses a gap in the research about the way that users experience content moderation actions by the online platforms.

At the heart of the paper is an examination of the colloquial concept of a 'shadow ban' to reveal ways in which algorithms at work on the Facebook platform have the effect of reducing distribution of content, with consequent stifling of users' speech. A shadow ban is a term that refers to specific scenario where users' content is hidden or deprioritised without informing them. The paper reveals how the Facebook shadow ban can suppress content distribution without specifically targeting it for removal, raising serious concerns for freedom of expression and transparency.

The platform was not concerned with a determination of the legality or harmfulness of the content. Instead, it was concerned with 'behaviour', or activities of the Page visible only through digital patterns in the underlying data. The action taken by the Facebook platform was to limit the distribution of the Page content through the News Feed system, which curates the content on the user's profile and is also an algorithm.

This paper has benefitted from a corpus of data that was gathered from 20 UK-based Facebook Pages. It includes data from Facebook Insights, as well as screenshots and a cache of emails. This evidence has made it possible to look under the bonnet in order to verify the users' experiences.

In the course of examining the evidence, it has been possible to verify that Facebook was imposing restrictions on content distribution from the Pages by means of a function known as a 'feature block'. This resulted in their posts not being shown in the News Feed that list the content on their 'fans' Facebook profiles. The block limited the distribution of the entire stream of posts from the Page, over a period of time, which was usually a week but sometimes longer. It ultimately meant that very few people saw the posts.

The users were often not informed that the block was being imposed, or the reason for it. When they did receive a notification, it was opaquely worded and they did not understand its meaning. Hence the term 'shadow ban' would seem to apply.

The 'shadow bans' were imposed on the basis of criteria related to 'behaviour' that could include inviting 'likes' or frequency of posting. It's a technique that is rooted in computer security and draws on methods used to tackle spammers and deceptive behaviour. It is positioned as being a more benign response than content removals or account suspensions, however, as this paper reveals, it is not a softer option from the user's perspective. In fact, it is quite draconian.

The paper was able to validate that these 'shadow bans' had tangible effects on the users and a metric was devised to measure this effect. The evidence from the affected Pages revealed how the 'shadow ban' created a sudden plunge in their reach (unique users) of between 93-99 per cent, which in turn led to a consequential loss of engagement and a downward trend in reach over time. The measurable drop in audience for the Page is not much different from completely unpublishing the Page. It has also been revealed that Facebook can selectively or partially block Page content, where some posts get normal reach and others get almost none at all. In some cases, the blocks and 'shadow bans' are imposed repeatedly, for extended periods. They were also imposed in tandem with brief

periods of unpublishing. The combined effects of these blocks have a longer term effect in reducing the reach attained by the Page.

This raises serious concerns about freedom of expression and transparency. The imposition of 'shadow bans' in the cases studied, appears to arbitrary. It would in all likelihood fail the test of the legality principle (Smith 2020), and interestingly, it is also unlikely to meet the standards now being set by its own Oversight Board.

There is a question arising about the balance of freedom of expression and security that should be addressed. This paper concludes with a recommendation for quality controls on Facebook's internal processes, potentially with a form of *triage* to identify genuine, lawful content that has been caught up in the security net. Overall, a much better understanding should be developed around the automated processes and algorithms that are used in content moderation.

Content moderation and automation

Content moderation is that term used to describe the process by which the platforms determine whether or not items of text, graphics, images, or video should be permitted and the subsequent action to remove, restrict or allow. It is an acceptable term for a practice that is effectively a form censorship (Langvardt, 2018: 1358). It raises difficult moral and ethical issues around the discretion being afforded to the platforms to determine who may speak and who may not. A system that is too *laissez faire* allows the bad actors to undermine, bully, threaten and divide, and yet a system that is too rigid or requires draconian, means that global takedowns will suppress speech on a scale previously unimaginable. As such, content moderation sets up a challenging tension with free speech rights. For policy-makers, getting the balance right is challenging to say the least.

The political context is such that the online platforms are under pressure to remove illegal content, and content that is deemed to be 'harmful'. The latter poses many challenges because the legal notion of what is 'harmful' is overly-broad and ill-defined (Smith 2019: 5.17). 'Harmful' content includes forms of threatening or bullying content, but also sexually-explicit or violent content, suicide, inciting hatred, or content concerning illegal activities and drugs, and the list continues to grow. However, some of this content can cause danger to individuals or to society, and so policy-makers seek its instant removal.

Content moderation is operationalised via private governance systems, also sometimes referred to as 'voluntary' schemes or self-regulation. With regard to copyright enforcement, it is done in 'co-operation' with other private interests. The public facing rules are expressed in the terms and conditions of service (Sander 2020:946), supported by rules and norms that users are expected to follow. Examples of these norms include Facebook's Community Standards, and the Twitter Rules. However, these public-facing rules tend to be opaquely written, and content moderation decisions often follow internal rules that are not publicly available.

The global operation of the large online platforms means that they have to manage these requirements at scale. In order to meet this challenge, they are becoming reliant on automated systems. These systems carry out functions like scanning and filtering databases of text, images, movies or music tracks. These checks may be carried out *ex ante* – prior to publication - before content is uploaded (sometimes dubbed the 'upload filter'). Alternatively, checks are carried out *ex pos*t – after publication, where automated scanners seek out content that has already been uploaded onto the platform as, for example, by Facebook seeking out terrorism content (Sander 2020: 946-7).

Automated systems for content moderation are a blunt instrument. They replace human intervention in the decision-making process. They have evolved over the past 10-12 years to make use of digital tools using algorithms, artificial intelligence and machine learning [United Nations 2020: 5]. Such tools were traditionally deployed for spam detection and for digital fingerprinting in for example, crime detection. Some techniques, have been patented. They notably include a patent registered by Facebook for a form of shadow-ban. (Justia.com 2019; see also Menegus 2019). The patent protects Facebook's ability to conduct automated moderation of comments beneath content posted on the Pages, by making the comment invisible to all, except the user who posted it.

These tools raise many challenges. Automated systems fail to address the root problems (Mozilla, 2020) especially where the harmful or illegal content is being disseminated by bad actors. Moreover, they are not good at recognising context (Windwehr and York 2020) and often fail to do so. This is a critical weakness, because context can make a significant difference as to whether content is legal, or appropriate, or harmful, or not, as highlighted in a recent decision by the Facebook Oversight Board (2021).

A specific matter for this paper is the use of automated systems to track 'behaviour'. This refers to activity on the platform that is mostly visible through the patterns in the underlying analytics data. For example, it could be the volume and frequency of posting or inviting 'likes'. This notion of behaviour derives from the cybersecurity field, where the online platforms face a challenge in how to tackle bad actors who deliberately use deceptive or manipulative techniques, at scale. Camille François identifies 'behaviour' as a key factor that the platforms look for in seeking to tackle disinformation campaigns, viral deception, hate speech, violence and extremism content (François 2019). Deceptive behaviour, she says, *" is a fundamental vector of disinformation campaigns*" that typically employ "*techniques to exaggerate the reach, virality and impact*" to mislead, and make themselves appear more influential than they are. Disinformation campaigns are often coordinated across different jurisdictions in order to hide the identity of the perpetrator or because they are deliberately seeking to interfere within another State.

The online platforms have been searching for scalable solutions to meet this challenge. Techniques similar to those used in spam detection are a mainstay of operations against this kind of behaviour. These techniques employ algorithms, which are pieces of code that carry a sequence of instructions designed to reach a conclusion or make a decision (United Nations General Assembly 2018).

Research into the way that users experience these content moderation practices suffers from a general lack of data. Apart from the studies around social media and elections, "empirical research on the broader impact of platform takedown decisions is rare"(Keller and Leersen 2020). There is statistical data on the volume of takedowns, but there is little in the way of quality data about individual decisions (Keller 2019). The available data is supplied by the platforms themselves in the form of 'transparency' reports. It therefore is difficult to assess what specific text, images or videos are being taken down or to find out how the platforms are applying the rules.

Users tend to reveal confusion about the process of content moderation and the criteria applied. The application of actions such as the 'shadow ban' tend to create uncertainty for users and raise suspicions about bias. (Suzor, Myers West et al 2019; Myers West, 2018). The public debate is replete with concerns about bias. As Keller (2020) points out, the voices of the political right complain that they are victims of bias, however complaints about being unfairly silenced come from across the political spectrum (Keller and Chang 2020).

Anecdotes tend to drive the debate. One such anecdote concerns the Italian satirical website Lercio. On 6 September 2017, it noticed a dramatic reduction in reach from 500,000 (peaking at 2-3 million), down to just 30 following the publication of a controversial article. Lercio used unofficial channels to enquire of Facebook what had happened, and was informed that it had been subject to new measures to tackle fake news. The article was subsequently restored. Lercio's experience is similar to experience of the Facebook Pages in this study, as will be seen.

Keller (2019) describes how she discovered the user's experience of having their content taken down, when her husband reported her own post. It was an image of a Rodin male nude sculpture. Within 2 hours of the report, she received a takedown notification on grounds that the post went against Community Standards on nudity. What's interesting about Keller's experience is that she was in a position to know that the takedown was the direct result of another user reporting it and how quickly Facebook had acted on the report. The users in this study had no such information, and were completely in the dark about whether Facebook's actions were the result of a user report or an automated decision.

Users' rights

Users are largely absent in the discourse around content moderation. When they do appear, it is usually as victims of 'harmful' content, rather than entities with positive speech rights. This paper hopes to make a positive contribution by examining the users' experience of content moderation through their own data.

In Europe, the right to freedom of expression is governed by Article 10 of the European Convention on Human Rights (ECHR), and is enshrined in UK law in the Human Rights Act 1998. It is a two-way right, to impart and to receive information. It applies regardless of frontiers, and it applies to publication online as well as offline. Article 10 enshrines the right to freedom of expression "*without interference from a public authority*". When the Convention was drafted, it was assumed that any interference with freedom of expression would come from the State, and not from private actors or large global monopolies. The circumstances where this right may be restricted are defined in Article 10(2), which says that any restriction must be prescribed by law, pursue a legitimate aim and necessary in a democratic society.

There has been little in the way litigation to challenge online platforms on freedom of expression. One exception is a current case in Poland: SIN v Facebook, where a Polish NGO challenged Facebook for unpublishing Pages. In July 2019, the court made a preliminary ruling that Facebook must leave online all of the live Pages, but did not rule on the takedowns. At the time of writing, the case was stalled. (Fundacja Panoptykon 2019)

Previously, there was challenge in the French courts (L'Origine du Monde'). The case concerned a photograph posted of a painting that hangs in the Musée d'Orsay. Facebook argued that it should be heard in the Californian courts (US jurisdiction), but the Paris court ruled it could be heard in France (EU jurisdiction). After an eight-year legal battle, it was resolved amicably in 2019 (Sutton 2019).

The lack of safeguards and due process has led some users to resort to seeking publicity or an intervention by a celebrity. This has been known to achieve the desired affect of getting restrictions removed. In one high profile case, Facebook blocked a famous photo from the Vietnam war - the image of the 'napalm girl'. The image was reinstated after the Norwegian Prime Minister intervened (The Guardian 2016).

When it comes to 'behaviour', the safeguarding of users free speech rights becomes more complex. François (p 4) argues that it is challenging to address these issues through traditional regulatory techniques, reminding us that algorithms, and platform recommender systems in particular, encourage precisely the kind of 'behaviours' that the behaviour monitoring is designed to reduce. The user does not know what that they have done wrong that the system has detected. Where a user has acted in good faith and unwittingly uses similar behaviours to, let's say, a spammer, it raises many question. How should they know what the system has identified and how they can avoid the sanction. How will they be notified? These are important questions that are flagged up through the cases examined in this paper.

Public policy

If users' rights are absent in the discourse around content moderation, they are also missing in the policy framework. European policy is governed by the provisions on intermediary liability in the E-commerce directive (Article 14), which establishes that online platforms are expected to remove illegal and unlawful content, as well as copyrighted content, if they have 'actual knowledge' of its existence on their platforms. The highly controversial EU Copyright Directive 2019, asks online platforms to check for and remove copyrighted content on a voluntary basis, if they are informed of its existence by rights holders (Article 17 – dubbed the 'upload filter').

In January 2021, the Terrorism Content Regulation, currently being processed in the European Parliament, includes an obligation for every Internet platform to block terrorist content reported by the police in under an hour. NGOs fear that this will lead to even heavier use of automation. (La Quadrature 2021). The legal framework in the EU is in the process of being revised with the recently announced Digital Services Act (European Commission 2020) which does include requirements for some form of transparency in order to safeguard the fundamental rights of users, but it is unclear at this stage what kind of teeth it will have.

The online platforms are being put under pressure for increasingly fast removal of illegal and harmful content – from 48 hours down to as little as 2 hours, and in some cases instant removal or account suspension. This is highly problematic. An instant suspension has the *de facto* outcome of giving any user the ability to get another user suspended without question or due process. It also forces a default position on the part of the platform to restrict the content, as they have neither time nor (often) expertise to determine the lawfulness or otherwise of the content. It leads to increasing automation as the platforms seek to meet these demands.

These policies are characterised by an absence of transparency and accountability (EPRS 2019). Disclosure is minimal, and at the platform's discretion. There is no mechanism to hold the platforms accountable for the impact of their actions and their algorithms are protected as industrial secrets. The consequence is neither citizens nor regulators can know the criteria for platform actions nor assess how the platform is performing. There are gaping holes, from a regulatory perspective, around the way that the automated systems are trained. The attempt to regulate 'harmful' content falls at the first fence because defining 'harmful' is well nigh impossible (Smith 2019; Francois 2019). Then there is the question of bias which is often raised by users, but difficult to prove.

Until there is an understanding of how platforms operate, what they do and do not do, it will be difficult to devise effective regulation. Greater transparency is needed, but if it is merely based around statistics on takedowns, it will be forever ineffective. The opacity benefits only the platform, allowing it hide behind ill-defined numbers, but for the user it provides no insights. A disclosure framework based around principles or rules might at least allow users to know how the algorithms operate.

Overall, the regulatory framework lacks a mechanism to address content moderation that is algorithmically driven. A purely content-based approach, based on old-fashioned broadcast models of communications systems, will not cut it when content is being disseminated, manipulated and curated by algorithms. It is especially problematic where non-content criteria such as 'behaviour' are applied, as is the case with the so-called 'shadow bans' outlined in this paper. A number of experts call for alternative models . Keller (2020) suggests that lawmakers must cast a wider net for policy-making in order to understand the real trade-offs involved. Myers West (2018) conceptualises an educational, rather than a punitive model for content moderation. François (2019) proposes a triangular structure based on three factors: content, behaviour and actors. Such a framework would provide greater flexibility for the assessment of the platform actions. A more meaningful form of transparency and due process is called for by Suzor, Myers West et al (2019) such that users are given clear explanations of the decisions that have affected them, including the rules they have contravened, the specific content that violated the rules, and how it was identified. They also call for a better understanding of platform processes and systems, a thought that segues into the research that is the subject of this paper.

Through the lens of the user

The experience of the user is often expressed in a confused and chaotic way, but the underlying data can tell us a lot about how the platforms are acting. This study draws on data from a small sample of 20 Facebook Pages, that had all experienced ghosting effects that - from their viewpoint – were a mystery and had been imposed without notice or explanation. Analysis of the data made it was possible to look 'under the bonnet' of the platform interfaces and examine in some detail what was occurring.

A Facebook Page is a profile intended to serve the needs of an organisation, business or community. It is sometimes known as a 'fan Page'. It will typically promote content or inform a target audience about a theme or collective interest. It operates differently from a personal profile of an individual. One important difference is that Pages cannot collect 'friends', but instead other users can 'like' the Page. These are the 'fans' and liking the Page is an indicator that they are happy to receive updates from the Page about new content. A Page can have several people who are responsible for it, and they may hold different roles, such as administrator, contributor or analyst. These people will access the 'back-end' of the Page after they have logged into their personal account.

The ghosting effects described by the Pages in the sample were monitored over a period from September 2019 to January 2021. The 'ghosting' effect meant that Page was live, and the administrator could post, but the reach of the Page was dropping dramatically. An alternative description was that their Page had been 'hidden' or 'banned' . The users said it felt like the Page traffic 'went off the edge of a cliff'.

There were other effects reported alongside the ghosting, and these included unpublishing (taking down or removal of the entire Page), and the selective hiding of posts that included links to other websites. Some Pages reported that, in addition to 'ghosting', the Page administrator was blocked from posting.

This 'ghosting' effect experienced by the Pages conforms to what is commonly referred to as a 'shadow ban'. This term is intended to describe a scenario where users' content is hidden or deprioritised without informing them (Menegus 2019; Suzor and Myers West, 2019). It could be seen as a benign form of enforcement. However, as we have identified it in this research, a Facebook 'shadow ban' has measurable effects and arguably, it is one of the most draconian restrictions that can be imposed.

This research has identified that the 'ghosting effect ' or shadow ban is a restriction imposed by Facebook that reduces the distribution of the Page content (as will be outlined below). This restriction operates by means of the News_Feed algorithm to reduce the visibility of posts from the Page that are shown in the content displayed on user's timelines. Using the data provided for each of the Pages by Facebook Insights, it was possible to develop a metric and it was calculated that the reach was reduced by as much as 93-99% - where 'Reach' means the number of people who have had content from the Page on their screen. There is a range in the figure due to the variation in the size of the Pages (measured by number of likes) and the fact that they operate independently, so each one is different. Effectively, this is a reduction in the audience for the Page content.

Page characteristics

The research analysed data from a sample of 20 Pages that were operated by individuals and civil society entities. The sample included an anti-racism Page, a left-wing Page, a satirical Page, a Page aimed at NHS workers, and a number of Pages operated by local community-based NGOs. The Page sizes (measured by number of 'likes') ranged from two thousand up to just over 200,000. They are small compared with the size of some commercial Pages, which may be anything up to 1 million likes, and they are extremely tiny compared with celebrity Pages in the UK top 10, such as the singer Adele with 62 million likes.

The Pages were characterised by their high reach, with posts frequently going viral, attaining a reach of many multiples of the number of followers. For most of the Pages, the ratio of reach to the number of 'lifetime likes' was greater than 1. Typically, this ratio was between 1.25 and 5. it confirmed that they were reaching more people than they had fans. In fact, they generated an audience reach that was, in some cases, multiples of their follower base, an astonishing achievement.

The Pages attracted high numbers of daily likes relative to the Page size. For example, one Page was adding 12,000 likes in a 60 day period. They tended to be enthusiastic posters, sometimes posting at a rate of around 15-20 posts per day. The highest found, as a one-off, was a Page that scheduled 37 posts on one particular day. They posted a lot because they were operating in a dynamic and febrile political context and they had a lot to say. This compares with an average of 1.5 posts a day on a typical commercial Page [Rabo 2020].

Engagement - a ratio used to assess the level of active followers of the Page – was also high. There was a consistently high pattern across all 20 Pages. Engagement is a measure of the active users on the Page at any one time. It is about the number of people who don't just look at the post, but who react by clicking like, or comment or share. It was calculated as a ratio of the total daily engaged users for the entire Page (see *Methodology)*, and a/ the audience reach or b/ lifetime likes. When calculated as a ratio of reach, the engagement was typically between 8 and 22 per cent, but there were instances where it was much higher When calculated as a ratio of lifetime 'likes'

the percentages ranged from 10 per cent to an astonishing 473 per cent. In other words, there were more people engaging with the Page content, than the number of followers.

Here are some specific examples from the data. One of the larger Pages with 201,000 likes achieved average reach of over 252,000, and there were 42,000 daily engaged users. The engagement rate was 17.4% on reach and 21% on likes. Another Page with lifetime likes of 57,892 likes, managed to attain an average reach of over 1 million in autumn 2019, and over 99,000 daily engaged users. This reflects an engagement rate on lifetime likes of 171%. A small local Page, in early 2020, with only 4,832 users at the time, was achieving an average reach of around 40,000 and achieving an engagement on lifetime likes of 92%. Another Page achieved an average reach of 157,000 on a base of 35,000 lifetime likes, getting engagement rates on lifetime likes of 53%. As these figures illustrate, the Pages typically attain higher reach and engagement than their size - as measured in lifetime likes – would suggest. Compared with industry averages they have over-achieved. The norm for engagement ratio on reach is typically below 1% and engagement on likes typically averages 20-28% [Rabo 2020]. For NGOs, a normal engagement on likes can be as low as 1.7%. These Pages therefore punched well above their weight relative to their size, in the sense that their data revealed statistical patterns that outperformed the norm. It is not clear whether these over-achieving results - which would be regarded as an amazing success in a commercial environment – has anything to do with the reasons why they were targeted for the restrictions.

Blocking effects and shadow bans

This ghosting effect, commonly known as a shadow ban, was confirmed by the data analysis. Typically, it showed in the data as a sudden reduction in the reach. Depending on how big the Page was, the drop could be quite sharp. For example, it could go overnight from hundreds of thousands down to just a dribble. On a graph, it showed as a deep, U-shaped valley in between a series of peaks and troughs, and the U-shape rested on the bottom axis.

The effect typically lasted for 7 days, but sometimes it was 14 days. It could be repeated and extended over a longer duration. The most severe instance was a case in 2019 that it lasted for as long as 56 days. In another case in 2020, a period of restrictions began with a shadow ban, that was followed by other limits on the Page, over a duration of four months (Case Study 2). There was also a case of Pages being unpublished and then reinstated, with shadow bans before and after (Case Study 1).

The evidence indicates that these shadow bans were, in fact, a restriction on distribution. One clue was provided in a notification that came to some (not all) of the affected Pages. It stated “Stories are not being shown in News Feed”. It was alluding to the fact that the posts from the Page were not being offered to the users who had liked the Page. Posts are usually distributed via the News Feed. In very simple terms, News Feed is the system that pushes the content out to users, and displays the posts in a list on their personal profile. In technical terms it is a recommender system (Leersen 2020) that algorithmically curates the posts, and ranks them in order of presentation to the user (Cotter, Cho and Rader, 2017).

From the perspective of the Page, when stories – or posts - were not being shown in News Feed, that meant they were not being shown to users. It follows that users did not see them, and this would explain the sudden drop in reach. This drop had a measurable impact (see below). There was also a parallel drop in ‘likes’ and a fall in engagement. Therefore, not showing stories in News Feed put a squeeze on the distribution of the Page content, creating a negative cycle, pushing downwards on the overall audience or readership or supporter base of the Page.

When fully understood, this is not a ban in the true sense. However, it is a restriction placed on the Page without the user’s knowledge, that has the effect of hiding its content from users and has a tangible impact. An important distinction of the Facebook shadow ban is that it is a restriction on distribution applied to all the posts on a Page, for a period of time, and sometimes for an extended duration. This is a crucial differentiator, because of the impact it has on the Page reach and the way it can squash the audience numbers.

Notifications, as described above, appeared on what seemed to be an arbitrary basis. They were received for some of the instances, notably in the autumn of 2019, but it was not consistent. The notices, when they came, would

usually say how long the restriction was scheduled to last, but they did not give prior warning of the restriction being imposed. The duration seemed to be arbitrary. In most cases it was seven days, but could be repeated or extended to last weeks or even months. The users found that it was difficult to appeal. One asked how you could appeal when you didn't know it was you were appealing. Within this particular sample, some of the Page administrators were persistent in seeking a human contact at Facebook. Email replies were received in the autumn of 2019, and in spring and autumn of 2020. In some cases they were helpful, in others less so (see below)).

Another effect noted was what appeared to be a partial restriction of distribution in News Feed. Some posts would get normal reach, while others would get only a dribble. For example, Pages whose posts typically get a reach of several thousand, would find some posts only getting a few dozen. In general the Pages did not receive a notification about this restriction, and the effect was monitored via the Insights dashboard. It was observed that posts getting hardly any reach were those that included links, yet those that maintained normal reach were basic 'memes' – a graphic image which might contain some text, but otherwise no explanation or link. It resulted in some posts inexplicably being restricted whereas posts with similar subject matter and illustration were distributed as normal in 2019, some Pages received the following: "*Your Page has been blocked from sharing links*" but they did not understand what this meant and its meaning has only become clear subsequently.

In one particularly severe case (see Case Study 2), a Page experienced this effect on a recurring basis over a three month period in the autumn of 2020. There was no notification, but it was verified via emails from Facebook staff (see below) that this effect – a partial, selective reduction in reach of posts – was the result of a block that have been imposed by Facebook. It appeared to be a partial implementation of the 'shadow ban' restrictions on distribution described above. The basis for imposing it was not given. It did not reduce the Page reach as drastically as the full shadow ban, but there was a drop of around a third in Page reach for the last quarter of 2020.

These restrictions are noticeably different in their impact from content moderation in the sense that is generally intended in the policy debate. It is usually assumed that a single post or image will be removed on the grounds that it is illegal or breaches a Community Standard set by Facebook under its Terms of Service, whereas these restrictions affected the entire stream of posts generated by a Page on an apparently arbitrary basis.

They also were distinct from other forms of blocking actions that were logged for this research. These other blocking actions included blocks on posting, and warnings or 'strikes' that could be added up to inform an automated decision for a stronger restriction (see Content removal and strikes). There were only two instances of copyright enforcement: the content was removed and the enforcement notices were not appealed. There does not appear to be any connection between the restrictive effects experienced and the copyright enforcement actions.

Unpublishing

In November 20129, eight Pages in the West Midlands (see Case Study 1) were simultaneously unpublished. This meant they were taken offline. The posts and the administrator's interface entirely disappeared. The Pages lost all their content and their analytics. The decision to unpublish was taken unilaterally by Facebook. The removal was temporary and all of the Pages were subsequently reinstated, although no explanation was given either for the unpublishing or the reinstatement. The period of time for which each Page was unpublished varied between two to seven days. An analysis of the Facebook Insights data for the Pages confirmed that they were getting no reach during those periods. Based on information given by the Page administrators, there was no allegation of unlawful activity or violation of Community Standards, and no warning.

The unpublishing was combined with multiple 'shadow bans' restricting distribution of their posts. These were also imposed unilaterally by Facebook, and without prior notice or reason given. The 'shadow bans' were applied before and after being unpublished, with no explanation given. The overall effect was to reduce the reach of these Pages by 95-97 per cent.

The content that they were sharing was lawful and did not breach Facebook's Community Standards. The Pages had high engagement rates of between 10-15 per cent (daily engaged users as a ratio of daily organic reach). The Page

administrators were told to stop posting so much, and it was suggested that if they did so, the restrictions would cease.

Algorithms on patrol

It has been possible to confirm that the so-called 'shadow ban' that squeezed the distribution of the Page content in News Feed, was in fact a deliberate block – alternatively described as a limit or restriction - imposed by Facebook. This confirmation was obtained using the cache of screenshots and emails made available by the Page administrators (see *Methodology*). It was confirmed that the blocks experienced by the Pages were the result of automated systems or AI activity. It was also confirmed that the blocks did not relate to the content on the Page, but were due to some other unspecified factor. Many of the emails referred to the 'behaviour' of the Page, but did not define what the 'behaviour' was. From what can be ascertained, 'behaviour' refers to activity by the Page that may under certain circumstances, appear suspicious, or may reflect characteristics normally associated with spammer or other bad actors. None of this, however, was transparent to the user.

To take an example, one email confirmed that this reason for the block "*doesn't have to be content but could be things like posting too regularly or inviting likes too quickly*" but it failed to specify precisely what the Pages had done to incur the restriction. Nor did it say what was meant by 'too quickly' or 'too regularly'. However, it did confirm the restrictions were imposed "*automatically by our AI as a result of activity undertaken by the Page*".

This information helps us to build an understanding of what the so-called shadow ban really is. We already understand that it is a restriction on distribution of the posts in the News Feed (that curates posts for users), and that it can be applied to all the posts on a Page, or just some of them, for a limited period of time or for an extended duration. The reference to artificial intelligence (AI) informs us that the restrictions are imposed by automated system and the notion of 'behaviour' suggests that the reasons for imposing them are related to something other than the content. The shadow ban then becomes a restriction imposed on distribution of the Page content, on the basis of monitored activity by the Page.

Limiting behaviour

The confirmation that Facebook was, in fact, implementing blocks or limits on these Pages came in a series of emails from October 2019, and also from the autumn of 2020, seen by the author. The emails were written to a Page administrator , from Facebook support teams (sometimes signed as 'concierge'). One of the emails from 2019 explained that there was a "*feed limit restriction*", and confirmed that this was a '*temporary reduction in the account's distribution in news feed. The content is still visible on your timeline but fewer people will see it in their news feed*". It mentioned by name the *news_feed* algorithm. The email stated that the Page was " *restricted for causing people to like or engage with it unintentionally in a misleading way*." When the Page administrator queried what exactly the Page had done, and asked for specific examples, the response was: " *These limits are usually triggered when someone is commenting too much, posting too much, adding too many followers, etc*." No specific content or behaviour was cited.

The internal name given by Facebook to these restrictions is 'feature blocks'. The word 'feature' refers to features of the Facebook platform, which could refer to, for example, the ability to invite likes. It was explained that a block had been " *placed correctly"*, that a " *block extension*" was also in place – in other words that a seven-day block had been extended to 14-day block. Another response stated that a block could be re-imposed if Pages begin posting or inviting likes too quickly after a ban is lifted: "*Most blocks are temporary, and theirs has actually been lifted, but once they are allowed to use this feature again, they may need to slow down or stop this behavior so they don't get blocked again*."

The same language appears to form part of the standard corporate response to users who queried these feature blocks. Similar expressions have been noted in replies from 2019, and a year later in 2020. Here are Facebook's responses drawn from a string of emails regarding a Page that was subjected to repeated blocks in autumn 2020: *"We have limits in place to prevent behavior that other people on Facebook may find disruptive. These limits are usually triggered when someone is adding too many friends outside of their network, sending repetitive messages, or posting too frequently in groups".* [28 September 2020]

Then*: "Unfortunately, there is nothing we can do to avoid these limits as they are temporary and must be placed. However, you can avoid this situation in the future; once the limit has been lifted you can slow down or stop this behavior so you don't get blocked again.*" [30 November 2020]

And: "*As explained before, there is nothing we can do to avoid these limits as they are temporary and must be placed. However, you can avoid this situation in the future; once the limit has been lifted you can slow down or stop this behavior so you don't get blocked again.*" [1 December 2020]

And finally: "*Unfortunately, there is no more information that we can provide you about this matter. As explained before, there is nothing we can do to avoid these limits as they are temporary and must be placed. However, you can avoid this situation in the future; once the limit has been lifted you can slow down or stop this behavior so you don't get blocked again.*" [1 December 2020]

Facebook's Terms and Conditions, at the time of this study, did not define a feature block or the terms under which it may be applied. The Terms mentioned this concept only in a general way, as here under 'harmful conduct' where it stated: "*If we learn of content or conduct like this, we will take appropriate action – for example,... removing content, blocking access to certain features..*"

Interestingly, an email confirmed that it is possible for staff to get blocks investigated and removed, although this was difficult apparently, as per the following explanation: *"I am working on both getting these Pages investigated + potentially restored.... This is actually quite more complicated than one would think, because every Page feature block may be different and a different team works on different types of blocks.* 'Potentially restored' suggests that Facebook had discretion over what is allowed back online. The last sentence is especially interesting. It suggests there are multiple ways of implementing these blocks, and multiple teams internally working on them. It raises questions about the internal decision-making processes, and serves to highlight the lack of transparency. The email did indicate that it might be possible for an internal investigation to happen, and that UK stakeholder engagement staff would be the appropriate interlocutors for UK Pages.

News Feed

We have therefore learned from the evidence that the effects felt by users are 'feature blocks', that the blocking affects the News Feed, and that a notification received said '*stories are not being shown in News Feed*'. We have also learned that the blocks are being imposed on grounds of the 'behaviour' of the Page, and are quite distinct from content violations.

Facebook's News Feed is the means by which Facebook disseminates content to users on the platform, and is the primary way that many users get information. They can also get information via other methods such as search, but In practice News Feed is highly influential. News Feed refers to the list of content that appears in the timeline on a user's profile, however, as described above, it is much more than just a list. It is technically classed as a recommender system (Leersen 2020). This is a system that selectively presents content to users and shapes the way they get information. It functions according to criteria that include 'signals' from the user's own behaviour such as likes, clicks, shares, follows, and content uploaded. Over time, it will shape the behaviour of users. Confusingly, it is also the name of an algorithm that encodes this process ([News_Feed]). It is notorious for its lack of transparency and some experts liken it to a 'black box' within which the logic of its operation is invisible (Leersen 2020).

News Feed does not send content in chronological order from friends and 'liked' Pages, as users may think. Instead, it employs complex algorithmic techniques to curate the feed, selecting and prioritising the posts according to what it calculates could be the users' preferences, and as alleged by some experts, to boost its own advertising revenues (Leersen 2020). The ability to channel posts into users' News Feed and timelines is a feature of the Facebook platform that can be turned on and off.

Given its importance for shaping what users receive, it has remarkably little transparency. The inputs, that include user content and behavioural data, are not public, and likewise the outputs, such as the recommendations it is

making. Critically, News Feed not only promotes content, but downranks it as well (Leersen 2020), but it is not well understood how such downranking might function.

In particular, there is no framework to address the removal of content on the basis of 'behaviour'. It is even less well understood how News Feed functions from the viewpoint of a content curator or publisher, which might be a news outlet in their own right, a civil society entity or individual user who is using Facebook to inform an audience. From the perspective of the Page, the value of Facebook News Feed is in the 'push' distribution mechanism that sends out posts to users. Without it, the Page depends on the users to proactively search for or tune in. Hence, News Feed can be crucial to the success of the Page.

The findings of this study suggest that News Feed is manipulated by algorithms in order to suppress posts that are being published by Pages. This explains the notification '*Stories are not being shown in News Feed*'. By not recommending them, News Feed ensures that most fans of a Page will not see its posts. Reach and engagement are consequently reduced, the number of daily 'Likes' is reduced, and posts do not go viral. This is when users notice the 'ghosting' effect that they call a 'shadow ban'. The Insights Data verifies that Lifetime Likes come mostly from posts in the News Feed and from users Liking on the Page Profile. Thus, a block on distributing the Posts in News Feed will cramp the growth of the Page, because it gets fewer users liking it and overall it is likely to lessen the influence of the Page.

Facebook security policies

Ghosting or 'shadow banning' seems to be a technique derived from the computer security field to quieten bot or spam activity (Suzor, Myers West et al 2019) and keep an eye on potential troublesome accounts. This is akin to tactic used by Facebook to restrict the activities on its platform of extremist organisations as reported by the Financial Times [Lee, 2020]. It is deemed to be less severe than taking down the account, and allows for monitoring. It forms part of an overall set of techniques that include down-ranking, blacklisting, whitelisting that all derive from a security context. It is likely that the user does not notice. The Page administrators in this research did notice because they were watching how their individual posts were doing, and saw something was wrong when posts gained little traction.

By way of explanation, the possibility that these restrictions based on 'behaviour' could be rooted in security policy has been suggested. Indeed, the language of 'behaviour' can be found in discussions of Facebook's security policy and it resonates strongly with the explanations given by Facebook to the Page administrators. A blog post published by the platform [Gleicher 2019a], outlines a policy of regularly seeking out what it terms 'co-ordinated inauthentic behaviour' (CIB), defined as coordinated efforts to manipulate public debate for a strategic goal where fake accounts are central to the operation" [Facebook 2020c]. Enforcement techniques include '*temporary restrictions, warnings, downranking or removal*'. The blog post [Gleicher 2019a] says: " *the actors behind these campaigns are using deceptive behaviors to conceal the identity of the organization behind a campaign, make the organization or its activity appear more popular or trustworthy than it is, or evade our enforcement efforts. That's why, when we take down information operations, we are taking action based on the behavior we see on our platform — not based on who the actors behind it are or what they say.*"

In December 2018, Facebook's head of cybersecurity policy, Nathaniel Gleicher, said in a video posted on the platform that the policy applies "*when groups of Pages or people work together to mislead others about what they are doing*". He explained that the platform was using technology to automatically detect and stop fake accounts. The policy is only concerned with users *'engaged in deceptive or violating behaviour*', regardless of the user, the content or the intention. He defined 'behaviour' as 'influence operations, spam and hacking'. 'Deceptive' means the deliberate misleading of other users, and may be via a network of fake accounts. Influence operations often involve foreign actors, including State actors, and cross-border activity. Gleicher clarified that when Facebook takes action to address 'behaviour', it is not looking at the nature of the content that the user might be sharing. Interestingly, at a London press conference on 7 November 2019 in the run up to the UK General Election, Mr Gleicher again emphasised this differentiation between content and behaviour (Facebook 2019b) from an enforcement perspective.

Where's the harm?

We have established that the ghosting effect - so-called shadow ban - is a restriction on distribution via the News Feed recommender system. This restriction has the effect of reducing the audience for the content posted by the Page. The Page stays online but is subjected to a limit imposed by Facebook that is known internally as a 'feature block', and the users are in many cases not informed. The techniques used by Facebook appear to be rooted in computer security methods. The criteria used to determine the restriction relate to activity on or by the Page – 'behaviour' – that is mainly visible through patterns in the data. Importantly the criteria for the restrictions do not relate to the content, and the lawfulness of the content is ignored. Facebook is able impose these blocks unilaterally.

A shadow ban is often considered 'benign' but is that really the case? It does lead to uncertainty among users, and there is often suspicion that they are being targeted by others as discussed in Suzor and Meyers West (2019) - something that was also confirmed in discussions and emails with the users in this study.

The metrics obtained in this study suggest a more serious impact of the so-called shadow ban in terms of the way it erodes the audience of the Page. This puts a greater onus on the need to safeguard users, and also calls into to question how the large online platforms can create the right balance between freedom of expression and the security and integrity of the platform.

Establishing a metric

In the cases where Pages were unpublished, the harm was clear because they lost their entire audience for the period of the unpublishing. However, for those Pages affected by the restrictions outlined in this paper, the harm is more complex to evaluate because it is about a change in the reach of the Page. Reach is defined as 'unique users' and it means "*the number of people who had any content from the Page or about the Page on their screen.*" Reach fluctuates from one day to the next, and those fluctuations must be taken into account. Evaluating the harm to the Page in this regard, is not about measuring the difference between two constants, but rather measuring the change between two averages.

The metric we were looking for was the average loss of reach attributable to the block that Facebook was imposing on the Page (the so-called shadow ban). It was calculated by taking an average of the reach for the duration of the block, and comparing it to an average reach for 28 days prior to the day that block was imposed. The same calculation was done for all 20 Pages, and the data was obtained via the Facebook Insights dashboard, in order to provide consistent reporting criteria for all Pages. It was Facebook's own data, and thus it was deemed to be reliable, and consistent with the information that Facebook itself would use. Using this metric, it was calculated that Page reach was suppressed by up to 93-99% of the normal traffic. This pattern was repeated across all of the 20 Pages in the sample. The variation in the figures allows for the different sizes of the Pages and fluctuation in reach. It was also identified that a majority of new likes come from stories in News Feed. It would therefore seem that the reduction in distribution set up a negative cycle, suppressing engagement and the acquisition of new likes (the mechanism to build the audience or follower base).

A partial restriction was identified where some posts got normal reach, and others only minimal reach. In the case we examined, the posts that included links were the ones that lost reach, and those without links had normal reach. It could be observed by checking post reach, where it was evident when the restrictions were starting and ending, Hence, an apparently innocuous restriction from the viewpoint of the platform, may have had a longer term impact in reducing the audience of the Page. Taken all together, these restrictions or feature blocks, also known as shadow bans, can suppress the voice of the Page as a whole, sometimes over very long periods of time.

In some cases, the extended restrictions meant a loss of audience for periods of several weeks. In these cases, the so-called shadow ban would be imposed for 7 days, and at the end of the 7-day period the user would be notified that the block on distribution was to be continued for a further 7 days. In one case in the autumn of 2019, an extended block was imposed for a total of 56 days (Case Study 2). In a separate case for the same Page in autumn 2020, the Page experienced a 28-day total 'shadow ban' and a further 3 months of partial restrictions, where links

were suppressed, but memes with no links were allowed normal distribution. A quarter by quarter comparison shows how this combination of restrictions reduced the overall reach by around a third.

These metrics and comparative data inform us that the restriction on distribution via News_Feed has a tangible, negative impact on the Pages. Arguably, reach is a proxy for an audience or readership in conventional media terms. These figures for the drop in reach do give rise to concerns that the Facebook shadow ban is a powerful tool that is capable of interference with freedom of expression.

Human rights

The data showed how the so-called shadow ban can result in a muffling or silencing of civil society voices and lawful speech, that occurs over extended periods, where the basis of the restriction is unrelated to the content. From the user's perspective, the application of restrictions appeared to be arbitrary. They were applied unilaterally by the platform, without warning. There was no allegation made that the user was acting unlawfully and the blocks were applied using undefined 'behavioural' criteria. It was confirmed in email communications that suggested that the lawfulness of the content was not relevant to the actions taken, and did confirm that the restrictions were applied to address unspecified behaviours. Expressed more simply, the criteria for the restrictions were not related to the content – what was being said – but rather they related to how much and how often the user was speaking. When challenged by the user, it was suggested that the user should speak less to avoid the restrictions.

All of this should ring alarm bells. Human rights standards insist that a user should be able to foresee the consequences of his or her actions where they are likely to incur a sanction, and to take steps to avoid that situation. Any rules restricting freedom of expression must be clear, precise and publicly available and any action taken must be the least intrusive to achieve the desired objective. The repeat imposition of the restrictions over an extended period of time, as happened with one of the Pages in this study, is especially problematic, as is the selective blocking of some posts and not others. As shown above, the effect was to quell the success of the Page in terms of its reach and engagement. If this were a matter of public law, the test of the legality principle would surely apply. [Smith 2020].

Interestingly, this was confirmed by the Facebook Oversight Board (2021) - the body set up by Facebook itself to make assessments of content moderation actions. In its decision ins Case Number 2020-005-FB-UA, the Board assessed Facebook's actions in removing content according to human rights standards under Article 19 of the International Covenant on Civil and Political Rights (ICCPR) and with specific reference to the United Nations Guiding Principles of Business and Human Rights. It's assessment made reference to the requirement for removals to meet standards of legality, necessity and proportionality, and to have a legitimate aim. The Board issued a reminder that moderation actions should not be perceived by the user as arbitrary. It called on Facebook to provide a notification of the reasons for removing content, including the specific rule (Community Standard) that is being used as the basis for the removal.

Facebook's Community Standards are written in a way that gives the platform a lot of wriggle room, but make little sense to the user. The Community Standards said little, if anything, about behavioural criteria and the explanations given by Facebook staff suggest that these criteria are extremely vague. The lack of public information meant that the users could not foresee how these restrictions on 'behaviour' would affect them. As has already been highlighted, what is 'too much' or 'too many'? Some people simply have a lot to say, and public debate is healthy. How should 'too much' to be determined? Is it a certain number per day, week or month?

The Page administrator's position is exacerbated by the lack of opportunity to appeal or obtain redress. Facebook does state that users will be offered an opportunity to appeal content removals however, it's not clear how this applies when the grounds for the restriction is 'behaviour'. Some of the Page administrators in this study did try Facebook's appeals process. The evidence in this study indicates that the users were not always offered an opportunity to appeal (and under Covid-19 policy the appeal option was taken away). It is not surprising, therefore,

that Facebook's own transparency data indicates a small number of appeals. When the users tried the support links on the platform, the responses they received, as quoted above, were often opaque.

Freedom of expression rights also apply to the 'fans' of the Page, those users who have liked it and are happy to receive content from it. Under European law, the right to freedom of expression is a two-way right to receive information as well as to impart it. In other words, a right to be informed as well as to speak. Where decisions are being made without any form of transparency, as here, people who seek to receive information will never know they are being limited (Husovec 2021). Hence, the 'fans' will never know that they are not getting the information they have requested.

Balancing speech versus security

The curious fact of these cases is that the content was lawful. Therefore it should enjoy the right to freedom of expression, which states that it should be free from interference. Yet Facebook acted to impose the shadow bans, without any regard for the lawfulness of the content. This fact is set against the understanding that Facebook was imposing the so-called shadow bans using algorithmic techniques rooted in security policy. Facebook stated that behaviour – activity such as volume of likes or frequency of posting - is a key criteria for security purposes, and that it would take precedence over the nature of the content.

When two rights are in conflict, a court of law would normally seek to find a balance. Where then is the balance between freedom of expression and security?

It's arguable that the 'shadow ban' - in seeking to protect the security and integrity of the platform – has placed a disproportionate restriction on lawful speech. The application of security techniques changes the game from content moderation to permission to speak. It is no longer a case of a single piece of content being removed because it does not comply with a Community Standard. It's whole streams of content that are being suppressed over extended periods of time. The restriction is actively preventing Pages from reaching their audience, and hence it also stunts their growth.

Facebook security is concerned about bad actors who seek to disrupt the communities on the platform, as well as the activities they engage in, such as spamming users or spreading disinformation. Spamming may involve the use of malicious code, fraudulent offers or illegal content. The intent may be to hack either the user's own equipment or their Facebook profile. Disinformation seeks to undermine public discourse, economic activity and democratic discourse. Those behind these activities may be acting alone, or they may be part of a highly organised operation, funded by a foreign State.

Did these Pages, high performing in terms of reach and engagement, and above average frequency of posting and likes, somehow fell into the net that was cast by the security policy? If so, it must be questioned why the quality controls were not in place to catch mistakes like this. In one case, there was a lengthy email exchange. Where was the process to escalate the enquiry and resolve it, rather than continue to impose repeat blocks over a three month period? It is understandable that if Facebook is monitoring bad actors, it would not want to give away information about what it doing. However, is it not possible to include a form of triage in the process, so that these innocent cases can be pulled out?

From a security perspective, curtailing the growth in reach of a Page is regarded as a good thing. However, If they were a business, the restrictions imposed on these Pages would be an anti-competitive measure. Small businesses on ecommerce platforms have been granted some protection in law against the platform acting anti-competitively. (European Parliament 2019). Should similar protection also be granted to Pages – profiles intended for civil society entities and special interests who act as a content curator or media not as an individual? The Digital Services Act, currently being processed in the European Parliament, would seem to be an opportunity to consider this.

The weakness in the public policy approach is its concentration on the volume and speed of illegal content taken down. A limited amount of data is made public by Facebook in its quarterly transparency reports. An inspection of

the data for Quarter Three 2020 (Facebook 2020e), the latest available at the time of writing, suggests that very little is reinstated compared with the amount taken down, and the number of appeals is also low. This is not surprising, based on the sample in this research.

For users, whether civil society or commercial, online platforms are a vital route to disseminating their message and arguably they are a public forum for debate and discussion. An understanding of the platform enforcement process and the way that decisions are taken is a vital step to safeguarding those public fora. The public policy approach would be strengthened by the inclusion of regulatory requirements mandating quality controls for the platform processes alongside a more comprehensive transparency requirement. Abstract and over-broad terms and volume-based statistics are failing to keep the platforms accountable. There needs to be a requirement for explanations of decisions to restrict content, or that have the effect of restricting content, to include the basis of the decision and the process by which it was made. Clear and accessible information should be provided to the public about the community standards and any other rules that are used by the platform as the basis for restricting content.

Conclusions

The evidence in this paper suggests that the platform's actions were a disproportionate interference with the rights of the users concerned. The Facebook shadow ban is a restriction on distribution of the posts on a Page on the basis of monitored activity (behaviour) by the Page. The restriction operates via News Feed (the system, driven by algorithms, that curates posts for users). It results in a whole stream of lawful content being suppressed for a period of time, without explanation. The lack of redress, or even opportunity to appeal the restriction, is an illustration of how the imbalance of power of the platform is skewed against the user.

Policy-makers are coming around to the notion that the platforms must be made accountable for restrictions they impose. As an example, the provisions in Article 12 of the European Union's proposed Digital Services Act (European Commission 2020) on the need to inform users about restrictions are noted. However, these provisions are subject to amendment and how effective they will eventually be, is open to question.

Whether they are acting as speakers or as recipients of information, users have positive rights that must be protected and they should not be treated as passive victims. This paper supports calls for improved transparency, but the obligations on the platforms must be set in line with human rights standards as outlined above. It should go without saying that users should be informed of the specific reason for any restriction placed on their content, whether it is a removal of an individual piece of content or a shadow ban or other restriction that affects entire streams of content over a period of time. They should be offered a clear process to appeal any restrictions with rights of redress. In tandem, there should be a requirement for quality controls on internal processes of online platforms. These should be monitored by any future regulator and appropriate metrics developed. Finally, an improved understanding is needed of the way that the content moderation systems operate and their impact on users, and this paper hopes to make some contribution in that regard.

Acknowledgements

The author wishes to acknowledge the kind support of Scientists for EU and Dr Mike Galsworthy in establishing this study. Moreover, this study could not have happened without the assistance (and persistence!) of the Page administrators. I am especially grateful to Adam Brady, Peter Packham, Syd Cottle, Beth Linton, Caroline Kuipers, Richard Hunt, Mark Cunliffe, Schlomo, Michael Brett, Ruth St John, James Dart, John Gaskell and Dan Old, and all who sent me their data and answered my questions. I would like to thank Joe McNamee for his thoughtful feedback, and Professor Robin Mansell and Professor Juliet Lodge for their support. I am also grateful to Dr Miranda Mowbray, Heather Burns, Alec Muffett and Mireia Fontbernat, for their time and their considered opinions, that have contributed to shaping the ideas in this paper.

Methodology

The aim of this research was to confirm the existence of the 'ghosting effect' and the impact on the users. In order to achieve this, it was necessary to validate the actions that Facebook was taking, and to find a metric that could establish the impact of the blocks on the Pages.

This paper is about blocks imposed by Facebook that have the effect of suppressing content uploaded by users, but which are imposed, as far as we know, for a reason that does not concern the legality of the content, and may also not concern its compliance with Facebook's content standards, but does concern the 'behaviours' of the Page. It is understood that, in some if not all, instances where Facebook is looking at 'behaviours', it does not have regard for the content or the intention of the user. Therefore, for this research, the methods concentrated on examining the underlying data and other evidence. The content was not examined in detail, other than to establish that the Pages were in general terms acting lawfully.

A corpus of documents was gathered. This corpus contained data from the Facebook Insights dashboard of the individual Pages, including downloaded Excel spreadsheets and screen shots; more screen shots from the Page Quality and Support Inboxes, and a cache of emails supplied by the Page administrators. Some data was gathered directly by the author who had been given analyst rights to four of the Pages for that sole purpose. Analyst rights allow access to the data, but do not allow posting. The author had no role in uploading content on any of these Pages. However, having the analyst rights did enable her to witness what was happening and to validate what the Page administrators were claiming.

A metric was devised to measure the change in reach that occurred under the restrictions. An average Reach was calculated for the 28 days prior to a block being imposed, and an average of the reach during the period of the block. 28 days was chosen because Facebook was using 7-day periods and multiples thereof, in implementing the blocks. The difference between the two averages was then calculated. It is this difference that is quoted as the drop in reach caused by the block – in the results across all 20 Pages in the sample is was typically a drop of 93 per cent or higher.

To understand the characteristic behaviour of the Pages, figures were calculated for engagement. These figures were calculated using the Facebook Insights data that was downloaded for each Page. It was decided not to use the figure for engagement that Facebook provides on the Insights dashboard because it is not stated how that figure is calculated. There are different ways to calculate engagement. It can be calculated for individual posts or for the Page as a whole, and it may done using either reach or likes data. For this project, engagement was calculated for the Page was a whole. The calculation used daily engaged users as a ratio of both total daily reach and lifetime likes. A ratio of reach to lifetime likes was also calculated.

To observe the patterns of Page traffic, charts were created from the downloaded data. It was decided not to use the charts that Facebook puts on the Insights dashboard, but to create bespoke charts, so that traffic patterns over different periods could be examined. It was useful to see patterns over normal period (with no blocks), and compare to periods before, during and after blocks. These outputs were correlated against the cache of screenshots and emails.

Glossary

Algorithm A sequence of instructions intended to solve a problem or to make a decision, implemented in computer code.

Behaviour Activity on the platform that is mostly – and sometimes only - visible through the patterns in the underlying analytics data.

Content Moderation The process by which the platforms determine whether or not items of text, graphics, images, or video should be permitted and the subsequent action to remove, restrict or allow.

Content Removal Taking content off the platform, including all data associated with it. Usually refers to an individual piece of content that may be a post, image, text or video.

Engagement A ratio to indicate the level of reaction (shares, likes, comments) to content

Facebook Insights Analytical data for a Facebook Page, provided by Facebook

Facebook Page A profile intended to serve the needs of an organisation, business or community, or it may be used to publish content on a specific theme. It is sometimes known as a 'fan Page**',** where the users who like it are 'fans'.

Facebook Profile A space on the Facebook platform where users can post personal information and interact with people whom they have accepted as friends**.**

Feature Block A block placed on a 'feature of the platform, where a feature could be the ability to invite likes, or post, or obtain distribution to other users via functions such as News Feed.

Lifetime Likes The cumulative number of likes gained by a Facebook Page since its launch.

News Feed The system used by Facebook to disseminate content on the platform. It selects, ranks and curates content shown to users, usually as list on their profile. It is the primary way that many users get information, News Feed is a recommender system. It is also the name of an algorithm that encodes this process ([News_Feed]. News Feed is sometimes used to refer to the list that the user sees.

Reach Defined by Facebook as unique users: the number of people who had any content from a Page or about the Page on their screen. It includes posts, advertisements, check-ins, social information from people who interact with the Page.

Shadow Ban A specific scenario where users' content is hidden, demoted or deprioritised without informing them.

Unpublishing Taking an entire Page off the platform, including all posts in the timeline, and associated data.

Bibliography

European Commission (2020) Proposal for a Regulation Of The European Parliament And Of The Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC https://eur-lex.europa.eu/legal-content/en/TXT/?qid=1608117147218&uri=COM%3A2020%3A825%3AFIN

European Parliamentary Research Service (EPRS) (2019) A governance framework for algorithmic accountability and transparency April 2019

European Parliament (2019) Regulation (EU) 2019/1150 Of The European Parliament And Of The Council of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services

Facebook (2019b) How Facebook has prepared for the 2019 UK General Election https://about.fb.com/news/2019/11/how-facebook-is-prepared-for-the-2019-uk-general-election/ Checked November 2020.

Facebook (2020a) July 2020: Coordinated Inauthentic Behavior Report https://about.fb.com/news/2020/09/august-2020-cib-report/ Checked December 2020

Facebook (2020b) Community Standards Enforcement Report November 2020 https://about.fb.com/news/2020/11/community-standards-enforcement-report-nov-2020/ Checked November 2020

Facebook Oversight Board (2021) Case Number 2020-005-FB-UA https://www.oversightboard.com/decision/FB-2RDRCAVQ/

François, Camille (2019) Actors, Behaviors, Content: A Disinformation ABC Transatlantic Working Group, https://www.ivir.nl/publicaties/download/ABC_Framework_2019_Sept_2019.pdf, checked 7 January 2021

Fundacja Panoptykon (2019) SIN vs FB(Eng) https://panoptykon.org/sinvsfacebook/en checked 18 February 2021

Gleicher, N (2018a) Removing bad actors on Facebook https://about.fb.com/news/2018/07/removing-bad-actors-on-facebook/ checked 7 January 2021

Gleicher, N (2018b) Coordinated inauthentic behavior explained https://about.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/ Checked 7 January 2021

Gleicher, N (2019a) How we respond to inauthentic behavior on our platforms: policy update - about Facebook https://about.fb.com/news/2019/10/inauthentic-behavior-policy-update/ Checked 20 November 2020.

Husovec, Martin(2021) Over-Blocking: When is the EU Legislator Responsible? https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3784149

Justia.com (2019) US Patent for Moderating content in an online forum Patent (Patent # 10,356,024 issued July 16, 2019) https://patents.justia.com/patent/10356024

Keller, Daphne (2020) Before the United States Senate Committee on the Judiciary, Subcommittee on Intellectual Property, Hearing on the Digital Millennium Copyright Act at 22: How Other Countries Are Handling Online Piracyhttps://www.judiciary.senate.gov/download/keller-testimony

Keller, Daphne (2019) That time my husband reported me to the Facebook police: a case study In BoingBoing, 27 September. https://boingboing.net/2019/09/27/that-time-my-husband-reported.html Checked January 2021

Keller, Daphne and Leersen, Paddy (2020) Facts and where to find them: Empirical research on Internet Platforms and Content Moderation SSRN:https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3504930 Checked November 2020

Keller, Daphne and Chang, Ailsa (2020) New Executive Order to Expose Social Media Companies To more Liability for Content NPR https://www.npr.org/2020/05/28/864410781/new-executive-order-to-expose-social-media-companies-to-more-liability-for-conte?t=1610021705912 Checked 7 January 2021

Cotter, Kelley, Cho, Janghee and Rader, Emilee (2017) Explaining the "News Feed" Algorithm - an analysis of the "News Feed FYI" blog https://dl.acm.org/doi/pdf/10.1145/3027063.3053114

La Quadrature du Net (LQDN) (2021) Terrorist Regulation : LIBE Committee votes for authoritarian censorship Media release 12 January 2021

Langvardt, Kyle (2018) Regulating Online Content Moderation in The Georgetown Law Journal, Vol 106: 1353

Leersen, Paddy (2020) The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems European Journal of Law and Technology https://hdl.handle.net/11245.1/1f6a1f14-c064-4ff2-8b55-c56a8f9ae896`

Menegus, Bryan (2019) Facebook Patents Shadowbanning https://gizmodo.com/facebook-patents-shadowbanning-1836411346 Checked 17 February 2021

Mozilla Foundation (2020) When content moderation hurts https://foundation.mozilla.org/en/blog/when-content-moderation-hurts/ Checked November 2020

Myers West, Sarah (2018) Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms, in New Media and Society, Volume: 20 issue: 11, page(s): 4366-4383

Rabo, Olga (2020) We Analyzed 2,810 Pages to Calculate Average Facebook Engagement Rate, Iconosquare 19 August 2020 https://blog.iconosquare.com/average-facebook-engagement-rate/ checked 13 January 2021

Sander, Barrie (2020) Freedom of Expression in the Age of Online Platforms, in: Fordham International Law Journal, Vol 43:4

Smith, Graham (2020) Online Harms and the Legality Principle, 20 June 2020 https://www.cyberleagle.com/search?q=+legality+principle checked 14 December 2020

Smith, Graham (2019) Online Harms White Paper - Response To Consultation, 28 June 2019 https://www.cyberleagle.com/2019/06/speech-is-not-tripping-hazard-response.html Checked 12 December 2020

Sutton, Benjamin (2019) Facebook and a French teacher settled their years-long lawsuit over Gustave Courbet's "L'Origine du monde." https://www.artsy.net/news/artsy-editorial-facebook-french-teacher-settled-years-long-lawsuit-gustave-courbets-lorigine-du-monde

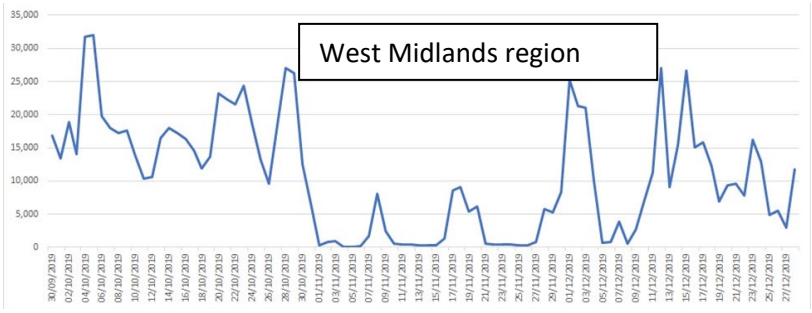
Suzor, Nicolas, and Myers West, Sarah, et al (2019) What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation, in International Journal of Communication 13(2019), 1526–1543

Reuters (2016) Facebook's Sheryl Sandberg on 'napalm girl' photo: we don't always get it right; In The Guardian 12 September https://www.theguardian.com/technology/2016/sep/12/facebook-mistake-napalm-girl-photo-sheryl-sandberg-apologizes

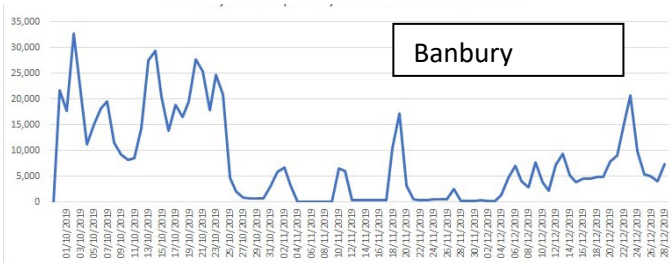
United Nations General Assembly (2018) A/73/348 Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression

Windwehr, Svea and York, Jillian (2020) Facebook's most recent transparency report demonstrates the pitfalls of automated content moderation https://www.eff.org/deeplinks/2020/10/facebooks-most-recent-transparency-report-demonstrates-pitfalls-automated-content Checked 30 December 2020

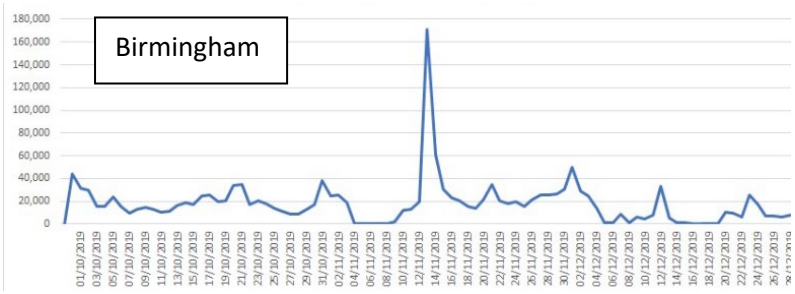
Case Study 1: Eight Facebook Pages simultaneously unpublished



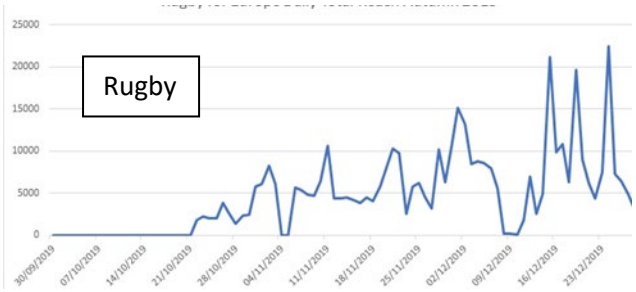
Eight Facebook Pages operated by civil society groups in the West Midlands were unilaterally unpublished by Facebook on 4 November 2019. The groups were located in Banbury, Rugby, Birmingham, Staffordshire, Stratford, Worcestershire as well as the Black Country, and there was one for the whole of the West Midlands. The combined average daily reach for seven of the Pages was more than 88,000. The period of time for which each Page was unpublished varies between two to seven days. An analysis of the *Facebook Insights* data for the Pages (see graphs) confirms that they were getting no reach during those periods.



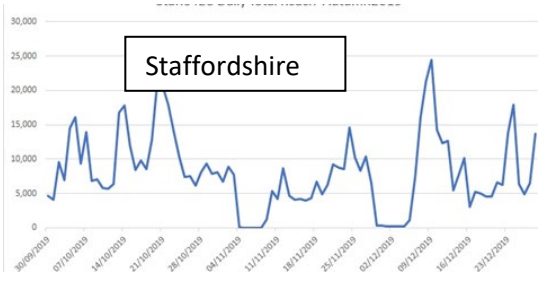
The unpublishing was combined with multiple shadow bans. These were imposed unilaterally by Facebook, and without prior notice or reason given. One was unpublished for two days in the middle of a shadow ban imposed on 1st November and removed on 7th November; the shadow ban was reinstated from 10th-16th November, 22nd-28th November, 5th-6th December and 8th -12th December. Another of the Pages experienced multiple shadow bans imposed during November 2019 immediately following the reinstatement of the Page, and another experienced shadow bans in December 2019. These restrictions had the effect of reducing the voice of the Pages by 95-97 per cent. The content that they were sharing was lawful and did not breach Facebook's Community Standards. The Pages had high engagement rates of between 10-15 per cent. (engaged users as a ratio of organic reach).



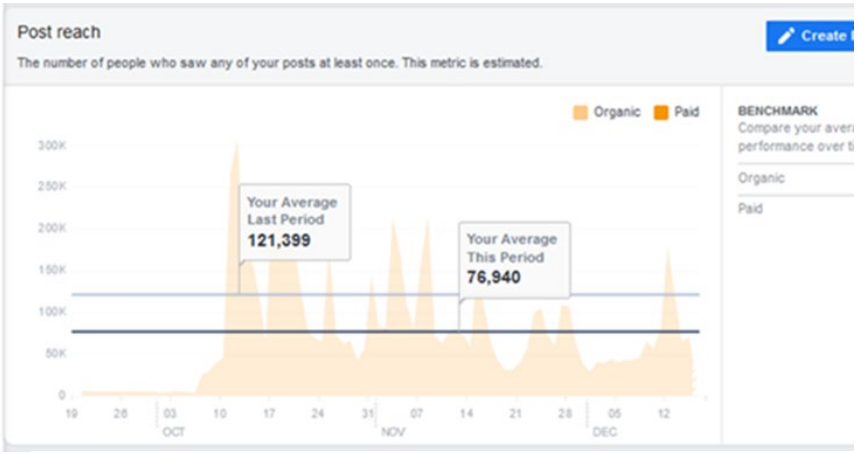
The Page administrators did not have a possibility to appeal the unpublishing, and they did not receive a specific explanation from Facebook as to why the Pages were unpublished. Contact was made on 6th November with a member of Facebook's UK staff who said that Facebook had changed its algorithms, and the Pages were advised to reduce the volume of posting to just 5-6 posts a day and not to cross-post or re-share.



This case raises questions around the accountability of the platform for actions taken through automated measures. The voices of people who were speaking lawfully were suppressed on a non-transparent basis that was not related to the nature of the content they were posting, and without a reason given except that they seem to be speaking too much. Arguably, it flags up a gap in the policy debate and raises serious questions for freedom of expression.



Case Study 2: Profile of a restricted Facebook Page



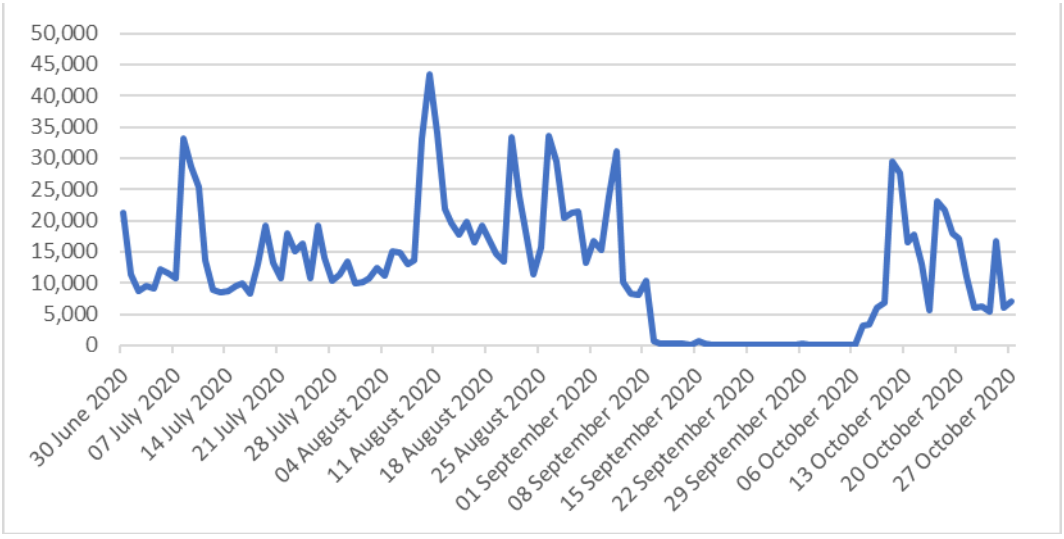
1 Comparative reach Q3 & Q4 2020. Shadow ban is on the left of the chart (September)

This Facebook Page was been subjected to ongoing restrictions for 119 days from 8 September 2020 to 4 January 2021. Facebook's own data reveals a drop in reach of around a third. This was the second time in 12 months that the Page had been restricted in this way. There was no specific explanation supplied, even though the Page administrator did manage to get in touch with one of the Facebook Concierge teams. There was no explanation given that identified a specific piece of content that was non-compliant with Community Standards. In emails to the Page administrator, Facebook did admit that it had placed restrictions on the Page, on the basis of unspecified 'behaviour' where this appears to mean activity that is only visible by examining data patterns, and could be something like posting 'too much'.

This case reveals how the Facebook platform can suppress lawful speech over extended periods of time without giving a specific reason. The sole basis for the restriction appears to an unspecified behaviour pattern. The means of restriction is the News_Feed algorithm that is coded to prevent posts from the Page being shown to its fans. The case raises very serious concerns for freedom of expression.

Case background

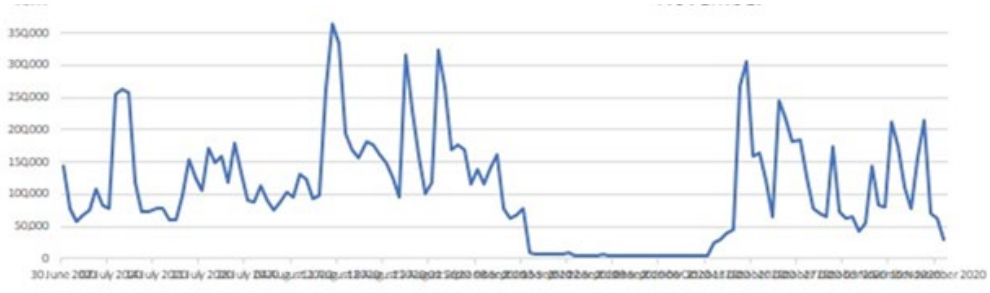
This is a public Facebook Page that curates content that is relevant to its followers. The Page has attracted 36,300 Likes, which is the standard measure of the audience size. The content is typically sourced from mainstream news organisations, and independent news outlets and original content. It includes a mix of articles, text, video, images, and memes. It frequently posts links to this content so that its followers can read the original source material. The content it shares is lawful and only content that is of interest to its target audience is shared.



2 Engagement June-October 2020. Shadow ban: 8 September - 6 October 2020

The relevance of the content it posts to its followers can be verified by the high engagement (see graph), and by the enormous reach that it is able to generate relative to its follower base (see graph). In commercial terms, this is a highly successful Page that engages its audience and provides them with information they are seeking. When assessed against standard criteria of audience reach and engagement it stands out. It is achieving levels of reach and engagement that are in double figures and very much higher than one would expect on a Page operated by commercial users. The difference is that it exists to serve civil society.

The shadow ban



3 Total daily Page reach June - October 2020: U-shape shows the shadow ban

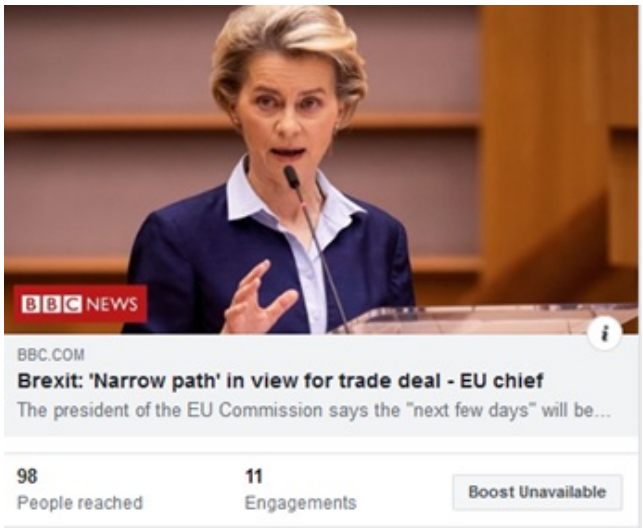
In autumn 2019, the Page was restricted 56 days from 8 October to 5 December 2019. In 2020, the Page suffered an extended period of restrictions from 8 September 2020 to 4 January 2021. From 8 September to 7 October 2020, the Page was subjected to a recurring shadow ban that reduced its Page reach from an average of 157,000 down to only 6,000, a drop of some 96 percent. There was no prior notification. There was also a block placed on the personal account of the Page administrator.

Selective blocking

Then from 16 – 31 October, the Page was subjected to another block. This one was different because it selectively blocked posts, rather than blocking all content. As before, there was no prior warning, nor explanation. A closer examination of the blocked posts indicated a common factor that they all contained an external link from the platform to another website. Posts that just contained an image or meme with no link did get reach.

It was again a recurring block, lifted an reimposed several times until it was finally removed on 4 January 2021. It had the effect of suppressing lawful content. For the quarter from 16 September to 19 December 2020, reach dropped by around a third, according to Facebook's own data.

Posts typically got thousands of views and good levels of shares and reactions, but under the block, some posts were logged with a reach of 10s. The content posted immediately prior to a block imposed on 27 November, got a reach of 55,984 and high engagement (2700 clicks, 3500 shares) . By 12 noon GMT the block was in place and the next post got a reach of 133, and negligible engagement with just 11 clicks and 6 shares. To take another example, a meme (no link) posted on 12 December 2020, when the block was in place, got a reach of 296,700. On the same day, the content linked to news articles got a reach of between 73 -146.



4 Top: normal reach 3,630. Below: blocked post - reach 98

Curiously, two very similar posts - shares of articles from the BBC website about the UK-EU negotiations, illustrated with a photo of European Commission chief Ursula von der Leyen – got different treatment, with different outcomes (see right). The top one, posted on 10 December 2021, got a reach of 3630, the bottom one, posted on 16 December 2021, got just 98. Both posts were about an important UK political story. The content is lawful. There is no issue with copyright on BBC content. Why was one suppressed?