Clean up the internet

April 2020

Time to take off their masks?

How tackling the misuse of anonymity on social media would improve online discourse and reduce abuse and misinformation

Contents

a. Introduction: How the current state of online discourse is harming our democracy

Computer-mediated communication, particularly on the large social media platforms owned by Twitter, Facebook and Alphabet/Google, plays a huge role in contemporary political discourse in the UK. The vast majority of MPs are active on Twitter and Facebook. For political journalists, social media is both where they often both get information and where they break their stories. For ordinary UK citizens, these platforms are a major source of political information, and a major forum for political discussion. They are, de facto, an important part of the contemporary public sphere and as such have a huge impact on the health of our democracy.

Yet it is widely accepted that the quality of discourse on these platforms is severely degraded by incivility, abuse and misinformation. This degradation takes several forms including:

- Threats, hate speech, harassment and incitement
- Legal but extremely unpleasant bullying, trolling, incivility and insults
- Misinformation and deception, both from non-organised individuals and organised networks

This poor state of online discourse is bad for democracy, both because it has a negative impact on the quality of debate and discussion it is possible to have, and because of its significant impact on who is excluded from such debate and discussion.

Incivility, insults and worst obstruct meaningful discussion, and impede understanding of different points of view. Whilst almost every social media user will encounter some level of unpleasantness, vulnerable and under-represented groups are disproportionately affected and at risk of being bullied out of political debates altogether. A 2017 Home Affairs select committee report noted abuse was targeted “particularly towards women and minority groups”. Research conducted by Amnesty International in the same year found that 1 in 3 women in the UK affected by online abuse reported having changed the way they express themselves online in response.

Misinformation further impedes sensible discussion of pressing issues, and fuels polarisation. Some of this misinformation is deliberately coordinated by networks of bad actors including foreign governments. It appears for example that in the wake of terror attacks in the UK, networks of twitter accounts purporting to be British but actually operating out of Russia, and apparently under the control of the Russian State, mobilised to introduce Islamophobic content into conversations. However, the effects of misinformation are widely felt and its dissemination is aided by the actions of a wider range of social media users. A 2019 study found that more than half of British social media users (57.7 percent) came across news in the past month on social media that they thought was not fully accurate, whilst 42.8 percent of British social media users who shared news stories were willing to admit to sharing inaccurate or false news.

b. The role of anonymity in incivility and abuse

There is no single explanation for why so many people so frequently behave so badly towards each other online. However there's broad consensus that the prevalence of anonymous, pseudonymous and unverified users on social media and other fora is one of the most significant factors contributing to this problem.

Anonymity has long been identified as a key factor in the Online Disinhibition Effect, which refers to the phenomenon of many internet users feeling able to exhibit behaviour which they would not exhibit offline. This sense of disinhibition can have benign consequences – making users feel able to challenge an injustice, or to explore ideas or emotions more freely, for example. However, the consequences can also frequently be toxic, with Online Disinhibition making users feel able to engage in negative online behaviours like bullying, harassment, and trolling.

One of the first academics to theorise the Online Disinhibition affect, psychologist John Suler, concluded in 2004 that "anonymity is one of the principle factors that creates the disinhibition effect. When people have the opportunity to separate their actions online from their in-person lifestyle and identity, they feel less vulnerable about self-disclosing and acting out."

Various studies have been conducted since, which confirm a link between anonymity, disinhibition, and various forms of toxic online behaviour:

- An Israeli study conducted in 2012 found that what they prefer to term an "online sense of unidentifiability" significantly increased the likelihood of "flaming" behaviour during computer-mediated discussions.
- A US study in 2014 found that newspaper online comment pages which permitted anonymous contributions had almost twice the levels of incivility of those which sought to require real names. The study authors concluded that "removing anonymity was a successful strategy for cutting down on the level on uncivil comment" and that "these findings should be of interest to those newspapers that allow anonymity and that have expressed frustration with rampant incivility and ad hominem attacks in their commenting forums".
- Another US study, from 2015, found that Twitter users to whom they gave an anonymous account were more likely to create and share sexist content than those given an account with personally identifying details. This same study also found that those users who initially shared sexist content under the cloak of anonymity, were then more likely to subsequently display sexist attitudes and behaviour offline.

These findings would likely resonate with other internet users, on the receiving end of uncivil or abusive communication online. A 2017 Pew Research Centre report found that in around half of all cases of online harassment, the victim felt unable to determine the real identity of the perpetrator. A 2019 report from the House of Commons Joint Committee on Human Rights noted that "Many of the MPs we heard from considered that anonymity fostered online abuse".

Clean up the Internet's own research into the attitudes of the British public also confirms that the link between anonymity and poor online behaviour is widely recognised. In polling which we commissioned, conducted by YouGov in February 2020, 83% of respondents said they thought the ability to post anonymously makes people ruder online

Lack of robust identity verification on social media platforms also makes it much harder for platforms to meaningfully enforce their existing terms of use or standards of behaviour. In the absence of any form of verification, it's harder to stop a banned user creating another account in order to send fresh abuse. More co-ordinated sources of extreme incivility, on the far-right for example, can maintain networks of accounts, and when one is banned for say, racist abuse of a politician, quickly deploy others with similar usernames and similar followers, to similar effect.

Of course, users' sense of anonymity, and the disinhibition and lack of accountability this leads to, does not fully explain all incivility or abusive behaviour in online forums. There are other common characteristics of computer mediated communication – lack of eye contact, for example – which psychologists believe also play a role. And there are prominent examples of out-and-proud trolls and bullies. But the weight of evidence does seem to suggest that anonymity is one of the main factors, and that were it to be addressed then levels of incivility and abuse could be expected to drop significantly.

c. The role of anonymity in misinformation

As in the cases of online abuse, harassment, and incivility, there is no single explanation for the prevalence of misinformation on the internet. Individual users spread rumours and falsehood, wittingly or unwittingly, for a variety of reasons and in a variety of ways. Deliberate and co-ordinated misinformation operations, such as those undertaken by foreign governments or extremist networks, make use of an ever-changing and diverse array of tactics.

However, the lack of options for proper identity verification on social media platforms, and the permissive approach towards anonymous and pseudonymous communication, make it easier for co-ordinated misinformation campaigns to operate. They also make it much harder for ordinary users to assess the reliability of sources and make properly informed judgements about what to believe, or what to amplify with a share or a retweet.

Analysis of Russian misinformation networks, such as those which targeted the US presidential election or those engaged in UK political debates, highlight a reliance on co-ordinated networks of inauthentic accounts which purport to be from users based in the target country, but are actually under the control of users based in Russia. In the absence of any form of rigorous identity verification, these networks of accounts are able to lend each other an appearance of authenticity and credibility by following and retweeting each other. Single users within such networks are able to use many different accounts, to reduce messaging loads and thus reduce the risk of being detected as malign, and to mitigate the impact of any individual account being shut down.

A detailed analysis of malign activity targeting Scottish political debates concluded that "4.25% of Scottish Twitter activity is identifiable as potentially malign" with "clear evidence of external botnets aimed at Scottish Twitter", with a particular targeting of controversial political subjects such as independence. This report identified anonymity and lack of verification as key weaknesses. To make it harder for external actors to pretend to be

Scottish residents, they recommend that Twitter users “be forced to show a confirmed geo-location (national only) in green, or a lack of confirmed geo-location in red”, and for Twitter to “introduce an option whereby a user can confirm their identity”. A recent study of

A detailed analysis conducted at George Washington University, of suspicious far- right German-language Facebook activity during the 2019 EU parliamentary elections, found that networks of inauthentic accounts remain a key tool in misinformation campaigns. The researchers identified tens of thousands of accounts with multiple suspicious features such as using stock actors’ images for their profile pictures; or changing their profile name multiple times; or using 2 character words which would not be recognised as names in Germany for both their first and last name; or having bought additional followers. They concluded that “a large network of suspicious accounts was active in promoting AfD Facebook pages in the lead up to the 2019 European Parliament elections”. This activity occurred despite Facebook claiming that it had clamped down on inauthentic activity, leading the EU Commissioner for Security to observe that “what platforms say and lived experience does not entirely align”.

It’s likely that anonymity, pseudonymity and lack of verification also increase the willingness of more ordinary users to amplify misinformation through sharing it. The 2019 study of UK social media users, by Loughborough University’s Centre for Online Civic Culture, found 17.3% of social media news sharers admit to sharing news they thought was made up when they shared it and 18.7 percent see upsetting others as an important motivation when they share news. For these users, the sharing of inaccurate stories appears to be yet another form of toxic online behaviour, analogous to trolling, likely to be fuelled by anonymity and its disinhibiting effects. Reductions in the levels of toxic online disinhibition, through restrictions on anonymity, could therefore be expected to reduce the prevalence of this behaviour.

d. Towards practical solutions: making the distinction between “benign disinhibition” and “toxic disinhibition”

Academic research frequently draws a distinction between “benign” and “toxic” forms of online disinhibition. Anonymity can play an important role in either form of disinhibition. Defenders of online anonymity often highlight, as an argument against seeking to tackle the negative effects of toxic disinhibition, the risks of also restricting the benefits of benign disinhibition. However, a proper understanding of the contexts in which benign and toxic disinhibition tend to operate suggests opportunities to restrict the *abuse* of anonymity, and its role in toxic behaviours, without sacrificing all of its potential benefits.

“Benign disinhibition” refers to the fact that an anonymous online environment can encourage users to feel more able to share information, emotions and ideas which they might feel inhibited from sharing with people they know and/or to whom they are identifiable. A whistle-blower, who feels able to sound the alarm about unethical activity by their employer thanks to their ability to tweet their revelations from an anonymous account would be an example of benign disinhibition. So too would be a LGBTQ+ teenager, concerned about the reaction of their immediate family and using an anonymous online forum to reach out for advice and support, a woman from a conservative religious background using a pseudonym on twitter to explore feminist ideas, or a trans person exploring different gender identities online.

Examples of “toxic online disinhibition” are sadly more universally experienced at present. An example would be the woman who knows her racist views are unacceptable and generally keeps them to herself, but feels able to use racist slurs and stereotypes as an anonymous user of a newspaper’s online comments pages. Or a misogynist who uses their pseudonymous twitter account to attack and threaten female politicians and journalists. Or a transphobe who hides behind anonymity to “deadname” a transperson. Or someone who’s got strong views about Brexit, is able to have fairly civil disagreements with his family and friends in person, but from behind the computer screen feels able to use inflammatory language and bombard those with whom he disagrees with exaggerations and insults.

A critical distinction between communication enabled by benign disinhibition and that enabled by toxic disinhibition is the difference between mutual, consensual exchange and unsolicited, unwelcome communication. The LGBTQ+ teenager is reaching out to people who want to engage with him. The whistle-blower is sharing information anonymously in the hope that others will value it and engage with it. By contrast, in cases of toxic disinhibition fuelled by anonymity, the communication is directed with negative intent at a recipient who hasn’t requested it. The racist or sexist troll sends abuse or threats to someone who doesn’t know them and that doesn’t want to receive it.

In general, users who wish to take advantage of the benign affects of anonymity want to communicate with users who actively choose to engage with them. In contrast, much of the worst communication fuelled by toxic inhibition is unsolicited and unwelcome. Were all users presented with a genuine choice as to whether or not to verify their own identity, and a genuine choice as to whether or not they wanted to hear from people who’d chosen to not verify their identity, this would therefore have a far greater impact on the dissemination of toxic content than it would on content which relies on anonymity for benign purposes

e. Towards practical solutions: making the distinction between authentic and inauthentic motives for anonymity

Similar distinctions are helpful when considering how to tackle the abuse of anonymous or false identity accounts for disinformation. There can be good reason for an anonymous account to be publishing information which they hope for others to disseminate further. A whistle-blower Twitter account such as The Secret Barrister (@BarristerSecret) is a good example of this – such social media accounts rely on anonymity in order to speak truth to power, and build up their credibility over time through the quality of the content they produce and the range of endorsements from identifiable figures who know the subject area and confirm its credibility. Crucially these accounts are up front about being anonymous and their reasons for doing so – it’s the *“**Secret*** Barrister”, a transparently anonymous voice, not a false identity.

Those who exploit anonymity in order to disseminate false information, on the other hand, rely on their lack of an authentic identity not being perceived by the recipients of their false information. Analysis of the activity of twitter accounts associated with the Russia-backed Internet Research Agency during the US presidential election in 2016 found extensive use of accounts, run out of Russia, which purported to be local news sources with handles like @OnlineMemphis and @TodayPittsburgh. Others purported to be local republican-leaning US citizens, with handles like @AmelieBaldwin and @LeroyLovesUSA, and yet others claimed to be members of the #BlackLivesMatter movement with handles such as @Blacktivist.

There are insufficient barriers on social media platforms to making false claims about identity or location. There are also no robust systems of verification available for users who make true claims of identity or location and wish to demonstrate that their identity is authentic. This creates an environment of indeterminacy in which it’s extremely easy for networks of inauthentic accounts to operate, and extremely hard for ordinary users to distinguish between authentic and inauthentic users.

If every user was able to see whether another user had chosen whether to verify their identity or not, and able to opt in or out of engaging with unverified users, this would make life much harder for those engaged in deliberate misinformation. Ordinary users would be empowered with new information about the verification status of other users, and apply their own judgement as to whether or not there were legitimate explanations for an account choosing not to be verified. A whistle-blower account issuing credible information over a sustained period of time, and endorsed by other verified users, would likely appear to have a credible reason for not being verified. So too would a user who makes it clear from their profile and their content that they are using a platform privately to explore, say, their sexual orientation or gender identity. It might raise more questions, on the other hand, if a local news-focused account claiming to be from Basildon had chosen not to verify that it was based in the UK.

f. Policy proposal 1: offer all users an option of verification

Social Media Companies should offer all users the option of verifying themselves through a robust and credible process, and making their “verified status” immediately visible to other users.

There are various ways in which this could be implemented. A variety of verification models already exist, which could be built on.

- Twitter used to offer a restricted availability, currently “paused”, account verification process (the “blue tick”). For the tiny proportion of users for whom it is available, the “blue tick” informs other users an account is authentic. Verified users are given additional filtering options, such as being able to see feeds containing only other verified users. Twitter stated that one reason for the suspension of the program in 2017 was a concern that verification was being confused with endorsement. Extending the option of a transparent verification process to all users, rather than offering it exclusively to those which Twitter has decided, through an opaque process, are “of public interest”, would address this problem.

- Facebook operates three different, and differently rigorous, verification processes - two levels of verification for pages, a “blue tick” and a lower level “grey tick”, and a separate and more robust verification process for political advertising on its platform, which includes verification of nationality.
- Several digital challenger banks such as Monzo operate identity verification processes robust enough to comply with EU Money Laundering Regulations. In the case of Monzo, verification processes have been applied successfully to over 2 million users.
- A number of mainstream dating sites such as Bumble have developed profile verification. Given the purpose of dating sites, verification tends to focus on confirming the authenticity of profile photos.

Another option which could be explored would be for tech platforms to partner with verification systems provided by a third party verification system. This could be provided by a government - an investigation into inauthentic activity in Scottish Twitter, commissioned by a SNP MEP, suggested that a “Scotland Verified” system could be piloted by the Scottish government. Several countries already offer some form of state-backed digital identity which would have the potential to underpin verification on social media, such as Estonia. Alternatively, it could be offered by private sector providers.

Any verification system would inevitably introduce additional steps which a user would need to take in order to achieve verification, and may require them to have access to some means of proving their identity. This would have the potential to raise some accessibility and inclusion issues. Our recommendation that verification be made optional would significantly mitigate these issues – no user would be at risk of losing access to a platform through not taking part in verification, whether by choice or because they had some difficulty with the process.

However, care would also need to be taken to ensure the process was as accessible as possible. The specific needs of different minority groups would need to be considered, for example to ensure people with no fixed abode were not excluded through not having a permanent address, or that there was a straightforward way for trans people to transition their accounts to their new name. The best way to ensure these needs are adequately considered would be to consult and involve a diverse range of users in the development of verification processes. Additionally there is extensive literature on best practice in this area which could be drawn on, and a range of charities serving communities with differing needs who could offer advice. A regulator could require social media companies to demonstrate that their systems are compliant with relevant equalities legislation, and that they have been developed with due regard to diversity and inclusion.

g. Policy proposal 2: offer all users an option of choosing whether or not unverified users are able to interact with them

Social Media Companies should offer all users a clear choice as to whether or not they want to hear from other users who have not been verified, and a user-friendly facility to filter out content from such users as a category.

This would have the advantage of empowering individual users to manage their communication preferences, without forcing those being targeted by anonymous trolls to play a game of perpetual “whack a mole” as they block their abusers one by one. So, for example, an MP who wishes to use Facebook and Twitter as a channel to engage constituents could continue to engage on the platform with those who have verified their identity and location, but anonymous trolls, who are possibly not even British citizens, would not be able to contact her with abuse or threats.

This solution would avoid seeking to place any new restrictions on the freedom of expression of those users who opted, for whatever reason, to remain anonymous. They would continue to be able to publish whatever content they chose, subject to the current restrictions (in theory at least) of the law of the land and the platform’s Terms of Use. Any user who chooses to do so would continue to be able to follow them, and to view and engage with their content.

h. Policy proposal 3: underpin these two requirements, by making social media companies who fail to implement measures to mitigate the negative effects of anonymity, such as those proposed above, legally liable for content produced by anonymous and unverified users

Social Media Companies and other internet publishers have long benefited from laws granting them immunity from liability as publishers of content produced or shared by their users. Whilst this relaxed legal framework has undoubtedly aided the growth of these companies, there is growing recognition in many jurisdictions that large internet companies need to be required to take more responsibility for the social impacts of the activities of their users, and that this requires some re-adjustments to this immunity.

Our own research supports the view that the public have lost patience with self-regulation of tech companies, and that there is strong support for changes to the law to force them to take more responsibility for abuse on their platforms. Opinion polling which we commissioned from YouGov, conducted in February 2020, found that 76% of the British public do not believe that social media companies are doing enough to protect users from anonymous abuse. A majority (52%, compared to only 17% against) believe that social media companies who do not take enough action to combat abuse should face criminal charges, whilst an even larger majority (80%) support large fines.

There’s a strong case that if large-scale platforms are to continue to enjoy the privilege of such wide-ranging immunity for the content they host and profit from, they should have to demonstrate that they have taken effective measures to limit the impact of toxic activity by anonymous and unverified users. A regulator should be able to independently review the steps a platform is taking to enable and encourage their users to take responsibility for the content themselves through offering robust identity verification. Where platforms choose not to take measures to make their users take responsibility for their actions, the platforms themselves should be forced to accept liability as a “publisher of last resort”. In practice, we’d expect this to act as a powerful incentive to large internet companies to act to mitigate the problems currently caused by abuse of anonymity and lack of verification.

i. Potential impact of implementing these measures

We would not expect the measures detailed above to act as a “magic bullet”. There are other factors which also contribute to toxic online behaviour, which our proposals would not address. Other interventions are also required – these might include, for example, more consistent enforcement of existing platform terms of use and existing laws regarding hate speech, and consideration of the ways in which content algorithms could be independently audited and required to reward quality over outrage.

However, there’s good reason to expect that these measures would have a significant effect. The academic research cited above suggests that when isolated as a factor, anonymity is a significant driver of toxic online behaviour and so measures to restrict it, even in the absence of any other changes, could have a significant effect. For example, the study of US newspaper comments pages mentioned above found that efforts to restrict anonymity reduced incivility by almost 50%, and concluded: “removing anonymity was a successful strategy for cutting down on the level on uncivil comment, [although] it by no means eliminated incivility altogether”. We’d hope that interventions to restrict anonymity would be considered as part of a broader, concerted effort to clean up online discourse, and might as a result contribute to a larger multiplier effect.

Additionally, whilst the measures suggested here are modest, nuanced, and informed by evidence, any new intervention or regulation can have unintended consequences. Therefore it would be important to monitor the impact over time, and make adjustments if unanticipated negative consequences emerge.

j. Conclusion

There’s compelling evidence that anonymity, pseudonymity, and a lack of identity verification are a significant factor behind the poor state of online discourse. The major platforms are currently failing to design their services in a way which restricts abuse of anonymity or the use of inauthentic identities. This has helped create a context where abuse and misinformation are rife, and where vulnerable groups are disproportionately badly affected. Given that so much political debate takes place on these platforms, the lack of quality and lack of inclusiveness in these forums are a threat to our democratic culture.

By drawing distinctions between the benign and toxic uses of anonymity, it is possible to start to identify modifications to platform design which would target the latter. If all users were given an option of verifying their identity, every social media user would immediately have a new, and significant piece of information to consider when assessing the reliability of another user’s content. Verified users would feel more accountable for their actions and less disinhibited, so would be less likely to misbehave. The option of filtering out unsolicited content from unverified users, and clear information as to whether or not a user is verified, would give users powerful new tools against trolling and misinformation.

We hope we have set out the urgent case for tackling anonymity, and laid out some credible options for how this could be done. We hope this paper stimulates debate as to the best way this can be done, and inspires others to develop further ideas to improve the state of online discourse.

j. Acknowledgements

This report was written by David Babbs of Clean up the Internet, with extensive input and advice from Stephen Kinsella.

We are extremely grateful for the feedback we received on earlier versions of this thinking from the following people: Professor Andrew Chadwick, Cori Crider, Mike Butcher, Emma Rose, Areeq Chowdhury, Professor Cristian Vaccari, Taniel Yusef, Alex Krasadomski Jones, Emeka Forbes, Chris Blackhurst, Robert McLeod, Dr Tamara Atkin, Jim Killock, James Morgan, Hannah Lownsbrough, Benedict Addis, Caroline Molloy, Lyra Dunseith, Seyi Akiwowo, Andrew Stroehlein, and Caroline Orr. They don't necessarily agree with all this report's conclusions, and any mistakes are the responsibility of the author alone.

All opinion poll figures, unless otherwise stated, are from YouGov Plc. Total sample size was 1,649 adults. Fieldwork was undertaken between 23rd - 24th February 2020. The survey was carried out online. The figures have been weighted and are representative of all GB adults (aged 18+).

The writing of this report and the accompanying public attitude research was funded by the Laura Kinsella Foundation.

Clean up the Internet is an independent, UK-based organisation. We are concerned about the degradation in online discourse and its implications for democracy.

We campaign for action from government the tech industry, to increase civility and respect online, and to reduce bullying, trolling, intimidation, and misinformation.

www.cleanuptheinternet.org.uk

contact@cleanuptheinternet.org.uk

07340 515932