ONLINE HATE SPEECH TARGETING THE ENGLAND WOMEN'S FOOTBALL TEAM DURING THE UEFA WOMEN'S EURO 2022

Arron Cullen and Matthew Williams

AUGUST 2022

HATE LAB

ANALYSIS HIGHLIGHTS

- The number of online hate speech posts targeting members of the England women's national football team was far outstripped by the number of hopeful posts (roughly one hate post for every 125 hope posts)
- 92 per cent of the team received targeted hate speech on social media platforms in the run-up to and during the tournament
- 20 per cent of players received over 50 per cent of the hate speech posts
- Hate speech was more prevalent on Twitter in comparison to Reddit or 4Chan
- 96 per cent of the hate speech identified was misogynistic, while 4 per cent was homophobic
- Many abusive posts used emojis to convey ridicule, disgust and sexist tropes, including the 'hot beverage' and 'nail polish' emojis
- There were temporal links between skilful football play and peaks in hope and hate posts
- Online hate speech tended to peak around England matches in the evenings between 7pm–10pm
- Public engagement with Twitter–based online hate speech was low, with most posts only receiving between 0-2 retweets or replies
- 91 per cent of the abusive Twitter posts were sent by users identifying as men[1]
- 54 per cent of users posting abuse on Twitter located themselves in the UK[2]
- Hotspots were found in London, Manchester, Liverpool, and Sheffield
- 52 per cent of the Twitter hate posting accounts were created in the past two years
- Only 3.6 per cent of the accounts posting hateful messages were suspected bots or fake accounts
- 94 per cent of identified hate speech remains live on social media platforms

92%
Received targeted hate speech on social media platforms

20%
Of players received over 50% of the hate posts

96%
Of hate posts were misogynistic

4%
Of hate posts were homophobic

91%
Of hate posts were sent by men

54%
Of hate posts were from the UK

3.6%
Of hate posting accounts were bots or fake

94%
Of hate posts remain live on platforms

[1] Of hate posting accounts where sex of the user (as indicated by name and/or photo) could be derived. Of total hate posting accounts, 61% were run buy users with a male name/photo, 6% were run by users with a female name/photo, and 33% could not be classified as any sex based on name and/or photo.

[2] Of hate posting accounts where location could be identified (as indicated by location metadata). Of total Twitter hate posting accounts, 28% claimed to be based in the United Kingdom, 24% claimed to be from other locations such as the North America, Europe, and Africa, and 48 per cent of hate posting accounts did not reveal their location.

SUMMARY

Following the Euro 2020 final, **HateLab reported** on the racist messages posted on social media targeting members of England's team. In anticipation of a similar response to the UEFA Women's Euro 2022, HateLab conducted real-time monitoring of online content posted in the build-up to and during the tournament (2nd May to 1st August 2022). This report presents an analysis of 78,141 social media posts directed toward individual members of the England women's national football team.

Using **award-winning** hate speech detection algorithms developed within HateLab, 380 posts were classified as either misogynistic or homophobic. While this report further evidences the continuing problem of online abuse directed at professional football players, it also highlights that hateful messages were outstripped by 50,422 posts that were classified as containing hopeful messages in support of the England team. This report builds upon HateLab's hate speech and divisive disinformation programme that supports government departments, law enforcement and civil society organisations across the world in their online monitoring and response activities.

THE STUDY

Previous **HateLab analysis** has revealed that professional football players have been receiving abuse online for over a decade. But it was the social media reaction to the on-pitch performances of Marcus Rashford, Jadon Sancho and Bukayo Saka during the Euro 2020 final that galvanised the attention online hateful abuse has received. The abuse of these players acted like a lightning rod for action. Following the final, arrests were made resulting in a few successful prosecutions, senior ministers met with social media companies, professional bodies and players to discuss the problem, and organisations including Kick It Out, Show Racism the Red Card, Stonewall (Rainbow Laces) and EE (Hope United) continued to raise awareness of hateful abuse impacting players and the community.[3]

EE's Hope United campaign highlighted the problem of sexist hate in the run-up to and during the 2022 UEFA Women's Euro Championship. Online misogyny is a pernicious social problem. Ofcom's 2022 Online Experiences Tracker found that women are more likely than men to be negatively affected by online hateful abuse.[4] Approximately 3 in 5 female respondents felt offended by trolling, compared to only 1 in 4 male respondents. The situation is worse for young women and girls. A 2021 Ofcom survey found that half of 12-15 year-olds reported encountering hateful content online, an increase on the 2016 figure of 34 per cent, with girls more likely than boys to report the content to platforms (33 versus 10 per cent).[5]

This study examined online hate speech that directly targeted members of the England women's team before and during the 2022 UEFA Women's Euro Championship. World-first technology, including the HateLab Dashboard and our unmatched **award-winning** machine learning algorithms that classify misogynistic and other hateful content in real-time, allow us to better understand the production and propagation of online abuse. Our analysis reveals that online misogynistic and homophobic abuse remains a problem in the women's game. Over 90 per cent of England's team was directly targeted with some form of hateful abuse around the time of the tournament. Twitter accounts created within the last two years that were run by users identifying as men located within the UK were most likely to send abuse. Abuse was most likely to be posted during 'flashpoints', such as key matches, including those where the women's team performed well. However, posts containing hopeful and positive messages massively outstripped the hateful posts during these flashpoints, and public engagement with Twitter based online hate speech was relatively low. Despite this encouraging finding, 94 per cent of identified hate speech remains live on social media platforms at the time of writing.

[3] Williams, M., The Science of Hate: How Prejudice Becomes Hate And What We Can Do To Stop It, London: Faber & Faber, 2022.
[4] Ofcom, 'Online Nation: 2022 Report', London: Ofcom, 2022.
[5] Ofcom, 'Children and Parents: Media Use and Attitudes', London: Ofcom, 2021.

KEY FINDINGS

HOPE

Over the course of the analysis period, England's players received 50,422 posts classified as expressing hopeful or positive sentiment.

Figure 1 shows that hope posts peaked when England played Norway, Spain, and Sweden. The largest spike occurred when the team won the championship. In-between matches, the rate of hopeful posts reduced to a consistent baseline level, similar to observed trends before the tournament began.

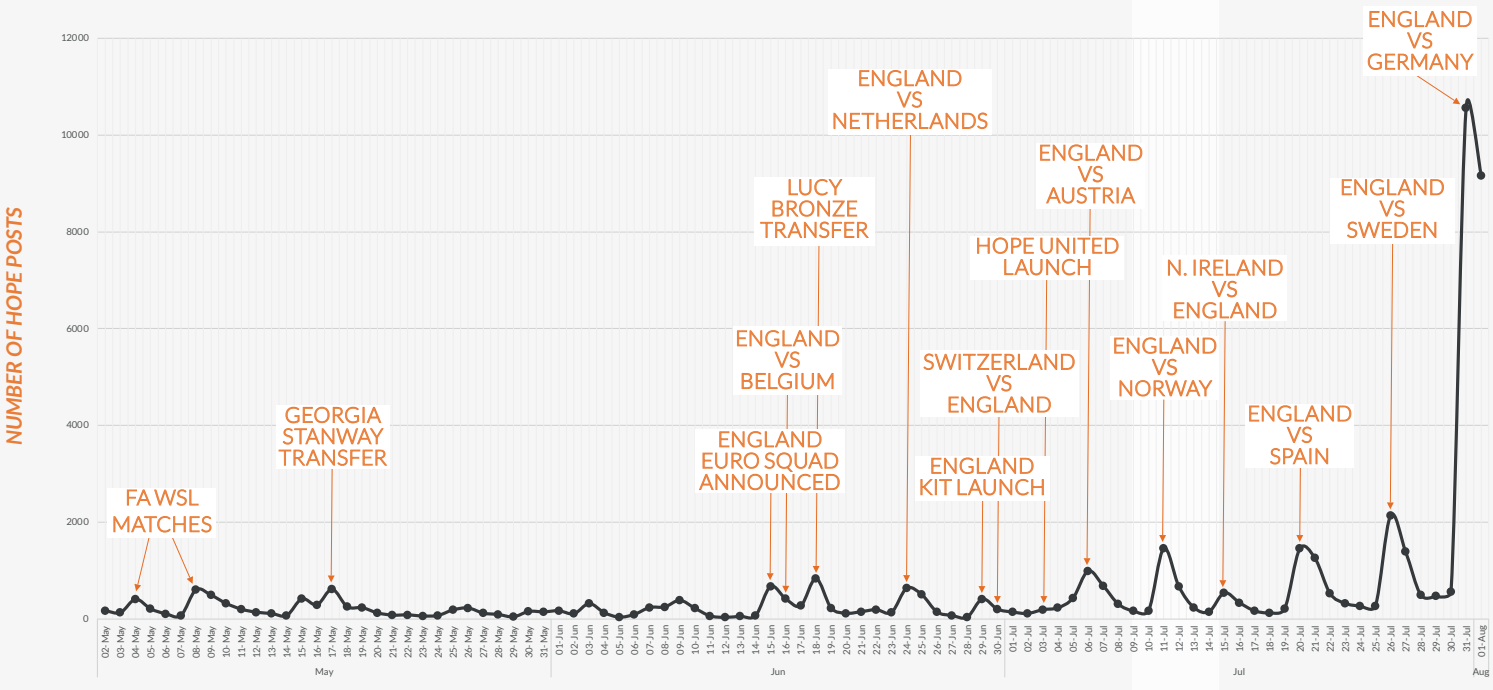

FIGURE 1

"HOPE FAR OUTSTRIPPED HATE."

Figure 2 shows that certain players, such as Beth Mead, Ella Toone and Leah Williamson, were more likely to receive hope posts throughout the monitoring period. The number of hope posts sent to all players far outstripped the number of hateful posts they received.

Research has shown that the portrayal of positive stereotypes by sportspeople, coupled with strong on-pitch performances, can suppress the expression of prejudice in fans. The 'Salah effect' resulted in a drop in online anti-Muslim hate speech from Liverpool fans, and a reduction in hate crime on the streets of Merseyside. The effect was down to Mo Salah's outstanding on-pitch performance in the 17-18 season, and his positive portrayal of Muslim identity. It is likely that the outstanding performance of the England women's team, and the positive portrayal of the skilful female football player, helped to break down negative stereotypes and prompted the expression of hopeful social media messaging by audiences.

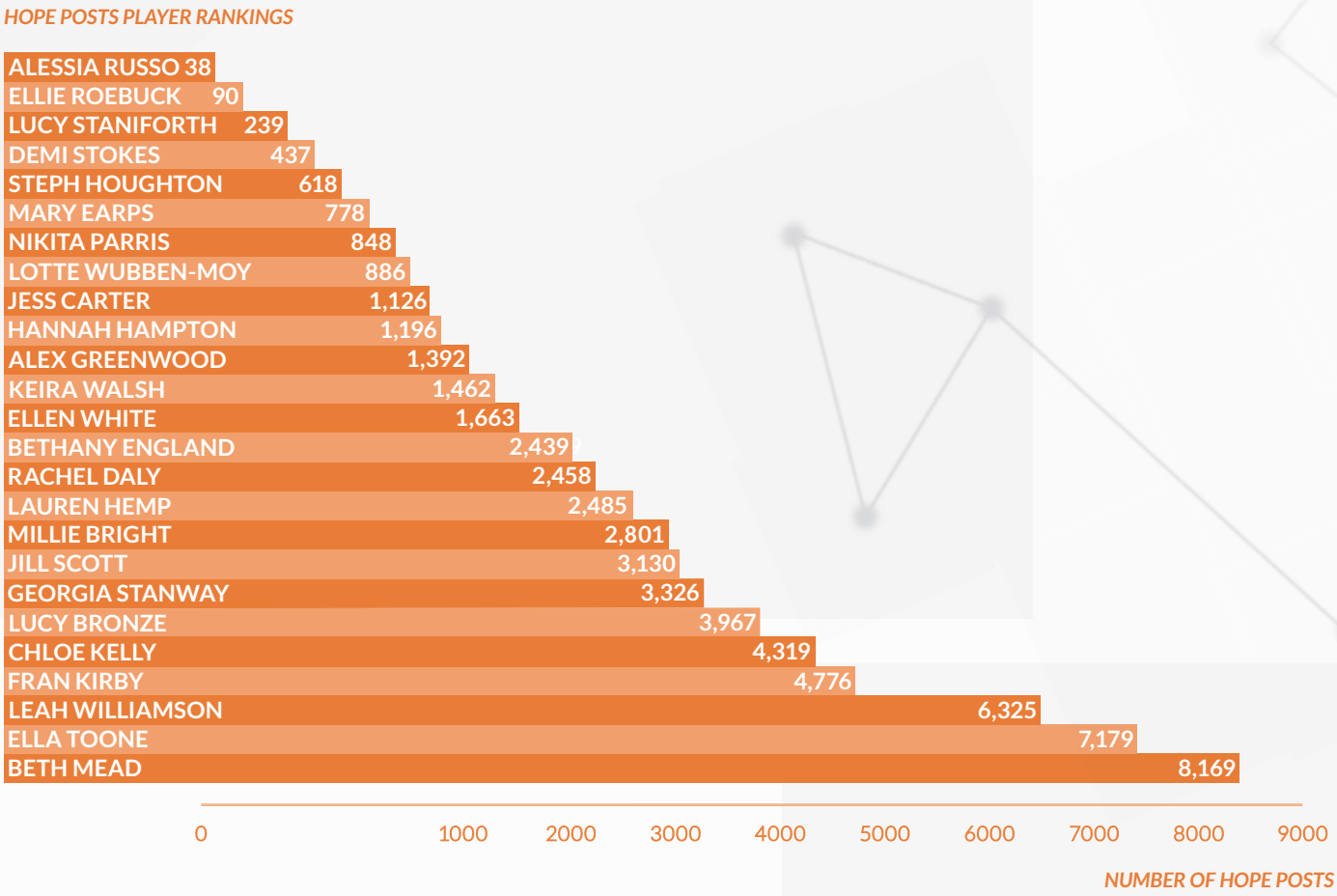

FIGURE 2

"ONLINE MISOGYNY IS A PERNICIOUS SOCIAL PROBLEM"

HATE

In total, 380 social media posts were identified as containing hate speech. Figure 3 shows that 97 per cent of hate speech was posted on Twitter, 3 per cent on Reddit and 1 per cent on 4Chan. Figure 4 shows that 96 per cent of posts were classed as misogynistic and 4 per cent as homophobic.

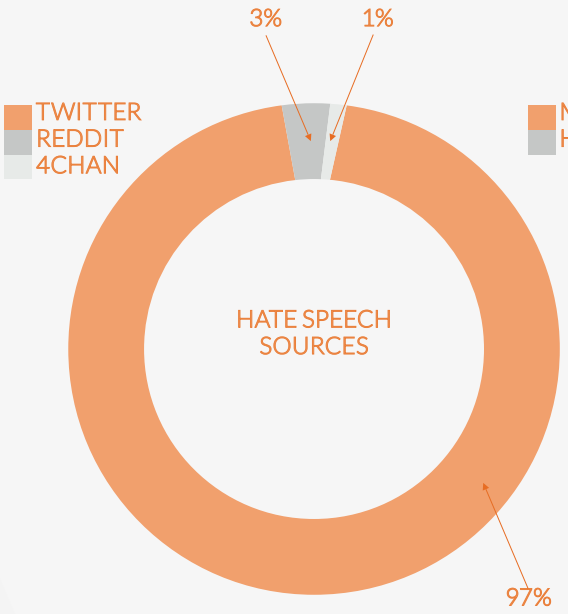

FIGURE 3

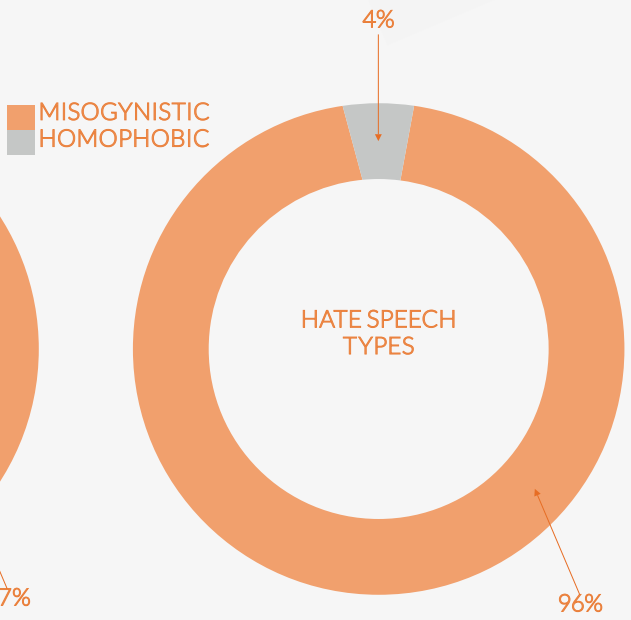

FIGURE 4

The posting of hate was more prevalent halfway through the 13-week analysis period, around the time the championship kicked off (Figure 5). The second highest peak was observed during the England vs Norway match (36 hate posts), where the score was 8-0 to England. The largest peak was during the final (93 hate posts) between England and Germany. Most of the posts used sexist slurs and tropes in their criticism of the quality of women's football. The final was televised live on BBC One in the UK, and it was widely reported that attendance at Wembley stadium set a new record, with 87,192 fans gathering at the venue the Sunday evening.

Hate posts tended to peak around England matches in the late evenings between 7 pm–10 pm, and then sharply decline to a consistent baseline rate between games. The finding that the posting of hateful abuse varied by key matches in the championship, with a sharp increase followed by an equally sharp decrease, is similar to findings in other research on the role of 'trigger events' in the expression of prejudice.

Several studies have found that in the aftermath of 'trigger events', such as terror attacks, court cases and political votes, a spike in hate crimes and speech was followed by a sharp de-escalation. This patterning suggests that event-motivated hate has a 'half-life'. [6]

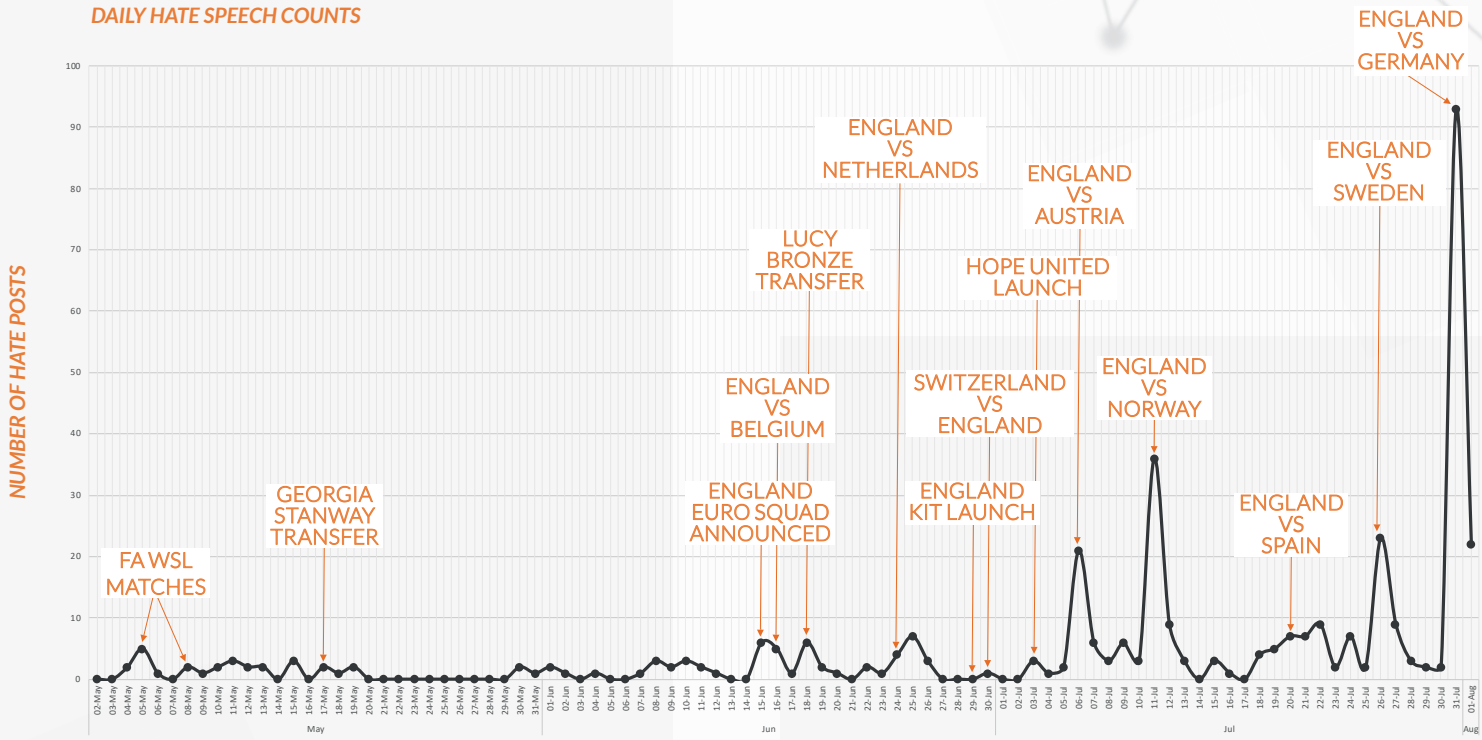

FIGURE 5

The idea that the expression of hate has a 'half-life' is based on the Justification–Suppression Model of the Expression and Experience of Prejudice.[7] The model assumes that we are all biased to some extent, but that we routinely control our prejudices due to suppression forces. The fact that it has become unacceptable to express prejudice in public, and the corresponding widespread desire to portray a positive self-image by appearing non-prejudiced, are both examples of a suppression force. The model also recognises the presence of justification forces that work against suppression. Increases in perceived threats to resources or identity from an 'outgroup' are both examples of a justification force.

Perception of threat can emerge from 'trigger-events' where an 'outgroup' is seen as attacking a sense of identity or taking a scarce resource. When justification forces are greater than suppression forces, some people are more likely to 'release' their prejudice in the form of hateful abuse.[8] After the 'trigger event' has unfolded and the sense of threat begins to wane, suppression forces kick back in resulting in a reduction in the desire to express prejudice. It is possible that some audience members observing the England team performing well in key Euros matches felt threatened in some way, which in turn acted as a temporary justification or 'releaser' of their misogyny and/or homophobia, in the form of an online hate speech post.

[6] Hanes, E. and Machin, S., 'Hate Crime in the Wake of Terror Attacks: Evidence from 7/7 and 9/11', Journal of Contemporary Criminal Justice, 30: 247–67, 2014. King, R. D. and Sutton, G. M., 'High Times for Hate Crimes: Explaining the Temporal Clustering of Hate Motivated Offending', Criminology, 51: 871–94, 2014. Legewie, J. 'Terrorist Events and Attitudes toward Immigrants: A Natural Experiment', American Journal of Sociology 118, 1199–1245, 2013. Williams, M. L. and Burnap, P.. 'Cyberhate on social media in the aftermath of Woolwich: a case study in computational criminology and big data'. British Journal of Criminology 56(2), pp. 211-238, 2016.

[7] Crandal, C.S. and Eshleman, A. 'A Justification–Suppression Model of the Expression and Experience of Prejudice', *Psychological Bulletin*, 129: 3, 414–446, 2003.

[8] Williams, M., The Science of Hate: How Prejudice Becomes Hate And What We Can Do To Stop It, London: Faber & Faber, 2022

Figure 6 shows the number of hate posts sent to each player, with 23 of the 25 England team members being targeted. Alessia Russo and Ellie Roebuck were the only players not to be sent hateful abuse via the @mention feature on Twitter or to be abused by name on Reddit or 4Chan.[9] The players who received the most hate speech across the analysis period were forward Beth Mead, striker Ellen White, midfielder Ella Toone, and defender Lucy Bronze. Around 20 per cent of players received over 50 per cent of the hate speech sent in the 13-week period.

It is possible that Beth Mead attracted significantly more hate speech posts than the other players because of her high profile and multiple successes during the tournament. She won the Golden Boot, Player of the Tournament, and Top Assist Provider, helping the England team win the championship and raise the profile of the women's game.

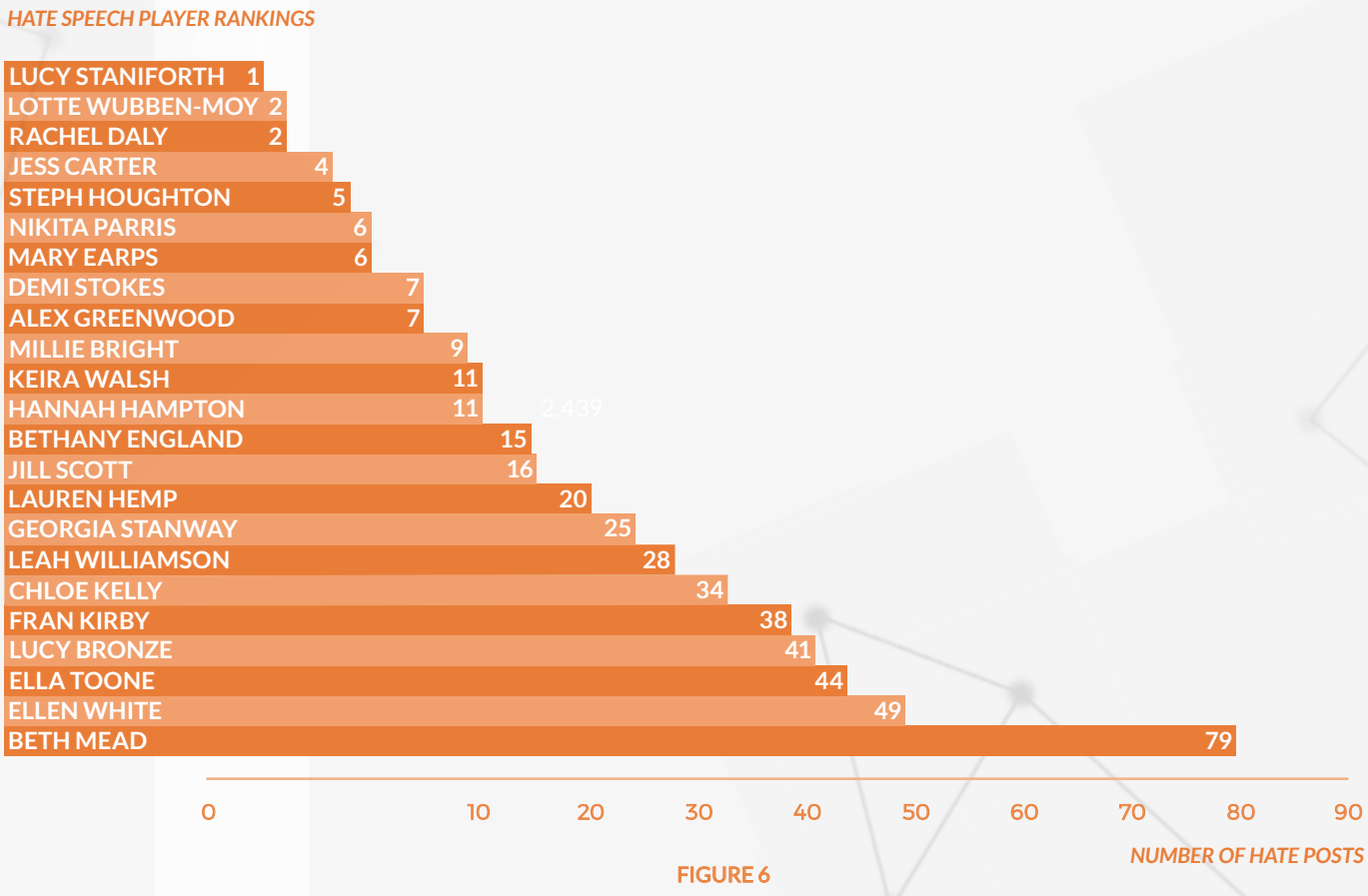

FIGURE 6

Previous research conducted by HateLab shows that online hate speech targeting male players is similarly not equally distributed, with high profile ethnic minority figures like Marcus Rashford and Mo Salah, attracting many more abusive posts than other players. While a high profile in the game can attract negativity, it is often accompanied by a significant amount of positivity, in the form of hopeful and congratulatory online messages. Beth Mead also received the most hopeful (positive) posts out of all team members during the tournament (see Figure 2).

[9] Players may have received abuse on other platforms or via private messages which were not accessible to the research team.

FIGURE 7

The word cloud in Figure 7 shows the most frequently used words (more frequent equals larger, less frequent equals smaller) across all misogynistic and homophobic hate speech posts. Most prevalent were terms related to criticising the quality of women's football. Some posts compared the women's game to Sunday league football, stating it was 'poor quality', 'rubbish', 'shite' and 'crap'. Many of these posts also contained sexist tropes and slurs, including the phrases 'get back in the kitchen' and 'make me a sandwich'. A number of these messages were posted in reaction to a point made by Gary Linker that made a favourable comparison to the men's game. A smaller number of posts questioned the sexual orientations of players in a fashion intended to ridicule.

Emojis used in hateful posts were also collected and analysed. Social media users frequently use emojis to convey emotion in posts. Brandwatch found that 95 per cent of internet users have used emojis in the past, and that over 10 billion are sent every day.[10] **HateLab research** has found that social media users are increasingly adopting the use of emojis to convey hatred in 'code' that can evade automated moderation systems used by platforms. For some time, Facebook, Instagram, and Twitter algorithms and moderation teams failed to recognise the use of certain emojis in certain contexts as potentially abusive. For example, the presence of a primate emoji in a post, while not always racist, can become read as racist by the receiver given a conductive context, such as a trigger event. Furthermore, the addition of emojis that convey disgust or anger, can hammer home the intended hateful meaning. HateLab machine learning algorithms recognise emojis in context, allowing us to classify hate speech more precisely and accurately.

FIGURE 8

The most frequently used emojis in the hateful posts directed at the women's team are presented in Figure 8. Laughing face emojis often accompanied hate posts, possibly to convey ridicule. Other frequently used emojis within hate posts included the middle finger as an insulting gesture, the vomit emoji to convey disgust, and the 'poop', trash and clown face emojis to portray the women's team as subpar, often in relation to the men's game. The hot beverage and nail polish emojis were also used alongside posts containing sexist slurs and tropes. The presence of a high number of emojis in hate posts is a concern given research that shows their use can increase arousal and emotionality over text alone.[11]

[10] https://www.brandwatch.com/blog/6-facts-about-emojis-found-using-new-analysis/

[11] Fischer B and Herbert C., Emoji as Affective Symbols: Affective Judgments of Emoji, Emoticons, and Human Faces Varying in Emotional Content. Front. Psychol. 12:645173, 2021.

HATE ACCOUNTS

There were 236 unique Twitter accounts sending hateful posts, with most users only posting abuse once. Only 9 users (3.8 per cent) posted 5 or more hateful posts, and the most prolific account sent 23 abusive messages. Metadata from these accounts was extracted to identify account creation date, sex of the user, location[12], tweet count and follower/followee count. Figure 9 shows that 123 hate posting accounts were created between 2020 and 2022. The remaining 113 were created between 2009 and 2019. This finding suggests that just over half of the hate posting Twitter accounts were fairly new to the platform.

There are several possible explanations for the high number of 'new' accounts sending hate speech. It could be that many of these new accounts are run by users who have had past accounts suspended due to breaches of Twitter's rules on posting hateful and abusive messages.

The users of these 'returner' accounts may therefore have a history of posting hateful content, and during the tournament may have relapsed into old behaviours.

Alternatively, some of these accounts could be run by newcomers to Twitter, meaning they are less familiar with platform rules on posting hateful speech, and therefore are less aware of the consequences, resulting in feelings of disinhibition leading to the posting of abusive messages.

It is also possible that some of these new users could be bots or fake accounts, created to promote discord and division. Bots are automated accounts that are programmed to retweet and post content for various reasons. Fake accounts are semi-automated, meaning they are routinely controlled by a human or group of humans, allowing for more complex interaction with other users, and more nuanced messaging in reaction to unfolding events. While not all bots and fake accounts are problematic (some retweet and post useful content), many have been created for more subversive reasons, such as influencing voter choice in the run-up to elections and spreading divisive content following national events.

Bots can sometimes be detected by the characteristics that distinguish them from human users. These characteristics include a high frequency of tweets and retweets (e.g. over fifty a day), activity at regular intervals (e.g. every five minutes), content that contains mainly retweets, the ratio of activity from a mobile device versus a desktop device, following many users but having few followers, and partial or unpopulated user details (photo, profile description, location, etc.). Fake accounts are harder to detect, given they are semi-controlled by humans, but telltale signs include accounts less than six months old and profile photos that can be found elsewhere on the internet which clearly do not belong to the account holder.

Only 3.6 per cent of the accounts posting hateful messages during the UEFA Women's Euros 2022 met the criteria for a bot or fake account.

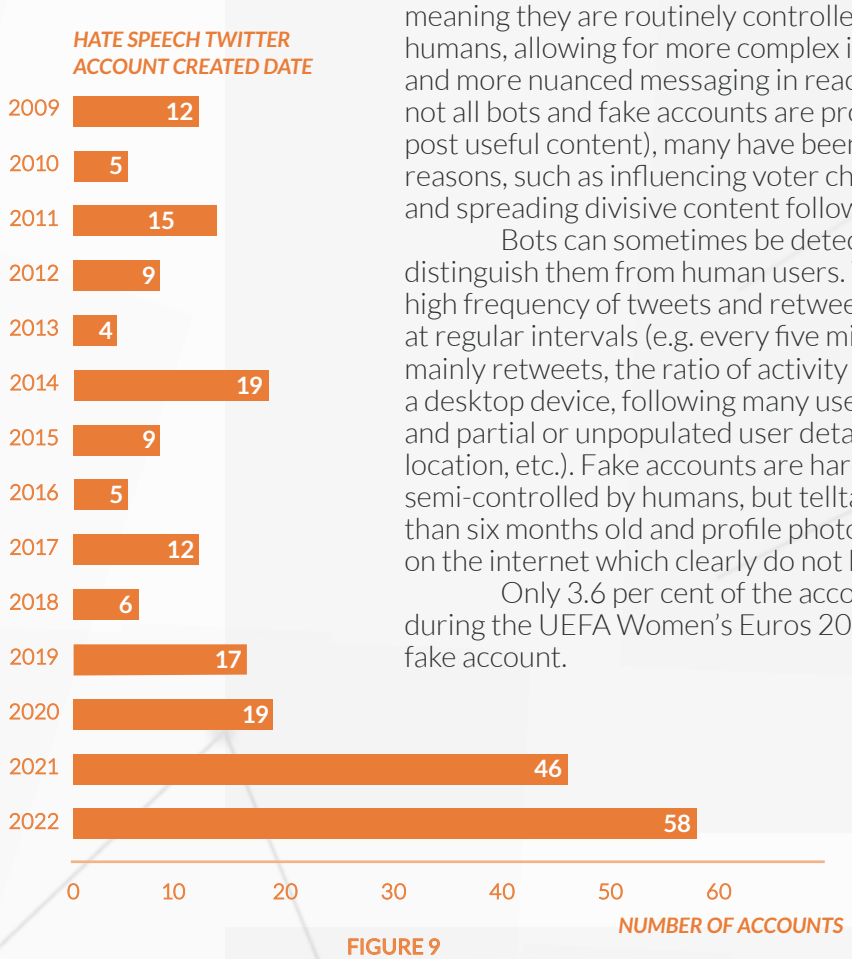

FIGURE 9

[12] Note that sex and location data were only available for 67% and 52% of accounts respectively.

Figure 10 shows that of the total number of Twitter hate posting accounts, 61 per cent were run by users with a male name/photo, 6 per cent were run by users with a female name/photo, and 33 per cent could not be classified as any sex based on name/photo. This translates to 91 per cent of hate posting Twitter accounts being run by men if we consider only those accounts where sex could be identified. This is not a surprising finding, and it reflects the offline experiences of women who are victimised because of their perceived gender. The Office for National Statistics (ONS) Crime Survey for England and Wales, shows that between 2017-2020, 89.4 per cent of gender-based hate crimes reported by women were perpetrated by men.[13]

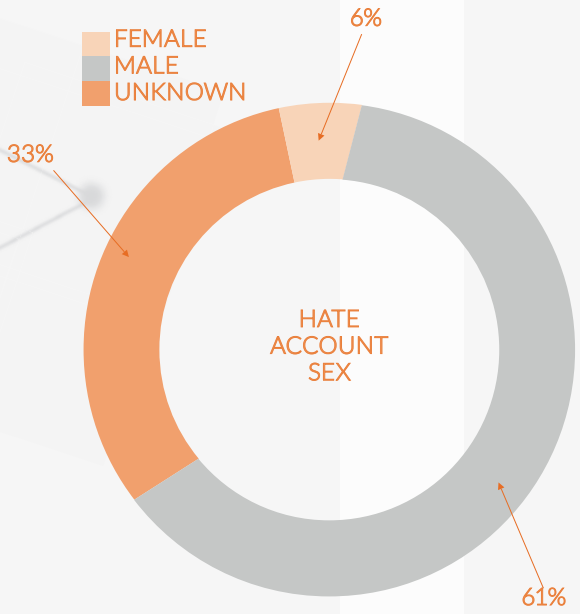

FIGURE 10

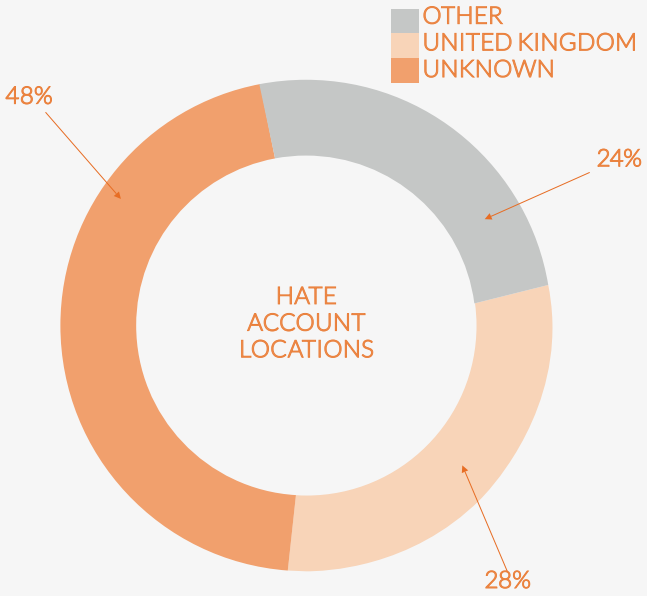

FIGURE 11

Of the total Twitter hate posting accounts, 28 per cent claimed to be based in the United Kingdom, 24 per cent claimed to be from other locations such as North America, Europe, and Africa, and 48 per cent did not reveal their location (Figure 11). This translates to 54 per cent of hate posting Twitter accounts being run from the UK if we consider only those accounts where the location could be identified.

[13] 4 per cent were perpetrated by females, and 6.6 percent were perpetrated by males and females.

Within the UK, hotspots were identified in London, Manchester, Liverpool and Sheffield, with smaller clusters in Newcastle, Birmingham, Bristol, and Cardiff (Figure 12).[14]

Previous **HateLab research** on hateful abuse in the Euro 2020 championship, and in the Premier League over the past decade, found similar results. Around 50 per cent of hateful posts sent to ethnic minority players during the Euro 2020 final, and around 40 per cent of posts sent between 2012-2021 to ethnic minority players in the current Premier League, originated from accounts claiming to be based in the UK.

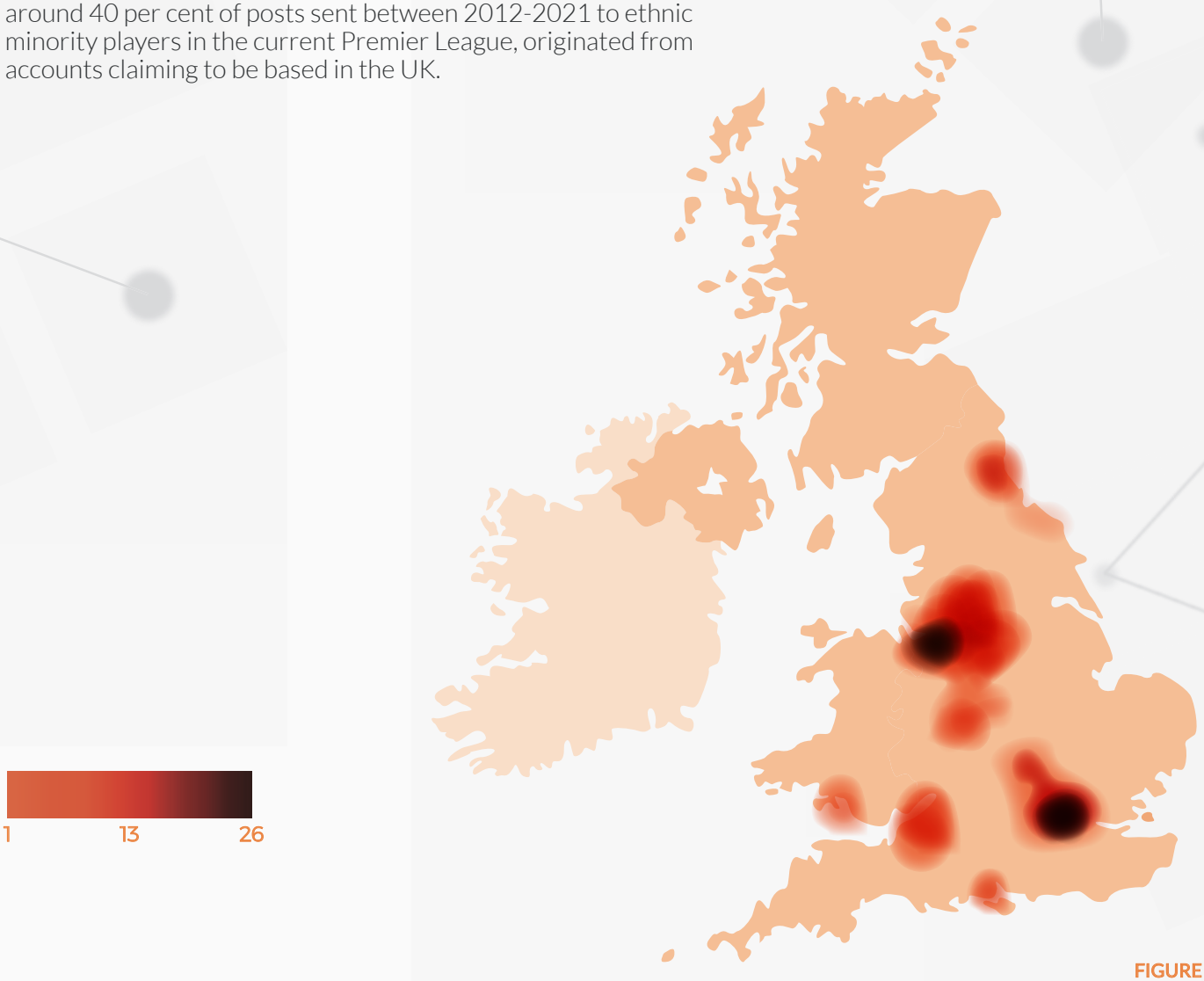

FIGURE 12

Combined, these findings suggest that a sizable proportion of perpetrators are operating within the jurisdiction of the policing services of England and Wales, Scotland and Northern Ireland. Where hateful abuse crosses the criminal threshold, it is possible, through cooperation with social media platforms, for formal law enforcement action to be taken. Where criminal abusers are locatable in countries outside of the UK, inter-agency arrangements across jurisdictions have made it possible for action to be taken by local law enforcement if suitable online abuse laws are in place.

[14] Not all hate speech posts included city level location information.

Metadata was extracted from hate posting accounts to analyse profile characteristics. Table 1 shows follower, following and tweet count data that suggests many hate accounts are well established, with the majority having several hundred followers and high tweet counts. The total number of followers in the hate posting network was approximately 100,000. While this seems large, it is important to note that the number of followers was not equally distributed throughout the network. A few users had extremely high follower counts (one with over 30,000), meaning a small group had a big influence over the whole network. This suggests that all hate posts did not have an equal chance of being seen by other Twitter users. However, as we only collected hateful posts that @mentioned players, it is likely that the intended receiver saw the post if they were using Twitter in the period shortly after England matches.[15]

HATE SPEECH TWITTER ACCOUNT INFORMATION

	M	SD	MIN	MAX	TOTAL
FOLLOWERS	422	2,124	0	30,151	99,668
FOLLOWING	474	795	0	5,000	111,758
TWEET COUNT	6,981	28,544	5	415,074	1,647,466

TABLE 1

Table 2 presents data on engagement with hate speech posts. These results indicate that public engagement is relatively low with most posts only receiving between 0-2 retweets or replies. The maximum retweet/quote retweet count was 5 for a single hate speech post, with a total of 24 across the network, indicating limited propagation (and possible endorsement). Replies, which consist of a mix of endorsements and counter-speech, were also limited, with one post attracting 20 replies, and a total reply count in the whole hate speech network of 266. Hate speech posts did receive more likes than retweets and replies, with one receiving 60 likes, and a total of 406 across the network, but again these are small numbers in comparison to overall traffic during the tournament.

The finding that hate speech posts sent during the UEFA Women's Euro Championship 2022 were limited in terms of their engagement and propagation resonates with previous research. Hate speech posted in the aftermath of terror attacks tends to be restricted to a limited audience of like-minded users, resulting in low propagation through the larger network and survival over time.[16] However, this finding should not detract attention from the fact that hate posts sent to members of the England team directly targeted their victims, meaning the intended receivers likely saw the posts, regardless of the overall low propagation and survival of hateful information flows.

HATE SPEECH TWITTER POST ENGAGEMENT

	M	SD	MIN	MAX	TOTAL
QUOTE COUNT	0.04	0.35	0	5	10
RETWEET COUNT	0.06	0.40	0	5	14
REPLY COUNT	1.13	2.80	0	20	266
LIKE COUNT	1.72	5.39	0	60	406

TABLE 2

[15] Some of the players were taking part in a social media blackout during the tournament, which may have decreased the likelihood that they would have seen posts via the @mention feature.

[16] Williams, M. L. and Burnap, P., Cyberhate on social media in the aftermath of Woolwich: a case study in computational criminology and big data. British Journal of Criminology 56(2), pp. 211-238, 2016.

"94% OF HATEFUL ABUSE REMAINS LIVE ON PLATFORMS"

CONTENT MODERATION

Hate speech posts were recollected after the analysis window to determine if they remained visible to footballers and the wider public. We found that 94 per cent of identified hate speech remained live on the social media platforms at the time of writing. This is far higher than the number of racist posts that remained online that targeted players in the Premier League (44 per cent) in the 20/21 season. It is also slightly higher than the number of abusive posts that remained online that targeted players in the Women's Super League (88 per cent) in the same season.[17]

Platforms' AI and moderation teams failed to pick up the majority of misogynistic and homophobic hate speech posts sent directly to members of the England women's team. Almost all of the posts picked up by our machine learning algorithms would be considered as either offensive or grossly offensive by the average person on the street due to their misogynistic or homophobic content (see methods). Furthermore, the targeted nature of the hate speech we identified arguably increases its severity and potential impact on the intended victims. Therefore, it is likely that existing AI solutions and moderation team processes are not working as intended to combat these forms of hateful abuse.

[17] Online Abuse: AI Research Study: Season 2020/21, Professional Footballers' Association, 2021.

THE IMPACTS OF ONLINE HATE SPEECH

Research on offline hate speech has found that victims experience trauma in a pattern that is similar to the response of victims of physical crimes. In some of the more extreme cases, the short- and long-term effects of hate speech are similar in form to the effects of burglary, domestic violence, assault and robbery.[18]

The reason for the extremity of harm from hate speech stems from the targeting of a person's core identity. Vilifying or dehumanising a person because of a fundamental part of their identity can generate negative emotional, attitudinal and behavioural changes. The impact is deeper if the victim is already vulnerable, for example suffering from depression, anxiety or the lack of a support network, and if the context in which the hate speech is uttered is conducive, such as where there exists a culture of fear, repression or intimidation.[19]

Short-term impacts lasting a few days can include feelings of shock, anger, isolation, resentment, embarrassment and shame. Long-term impacts, lasting months or years, can include low self-esteem, the development of a defensive attitude and prejudices against the hate speaker's group, concealment of identity and heightened awareness of difference. Remembering hate speech attacks has also been associated with increases in self-rated stress levels, and increased levels of the stress hormone cortisol in LGBTQ+ victims.[20]

Those without the cognitive tools to deal with online victimisation, namely the young, feel the effects most. A survey of over 1,500 young people in the UK found that those who encountered online hate reported feeling anger, sadness and shock. As many as three-quarters said it made them modify their online behaviour, including posting fewer messages or avoiding social media altogether.[21] Another study found young respondents most frequently exposed to online hate were less satisfied with their lives.[22] The average age of the England women's football team playing in the Euros was 27.

Online hate speech has the potential to inflict more harm than some physical acts, due to several unique factors. The anonymity offered by the internet means offenders are likely to produce more hate speech, the character of which is more serious because of the lack of inhibition. The temporal and geographical reach of the internet means hate has become a 24/7 phenomenon. For many, especially young people, communicating with others online is now a routine part of everyday life, and simply turning off the computer or mobile phone is not an option, even if they are being targeted with hate. Online hate speech then has the insidious power to enter the traditional safe haven of the home, generating a cycle of victimisation that is difficult to break.

[18] L. Leets, 'Experiencing Hate Speech: Perceptions and Responses to Anti-Semitism and Anti-Gay Speech', *Journal of Social Issues 58 (2002), 341–61.*

[19] Williams, M., The Science of Hate: How Prejudice Becomes Hate And What We Can Do To Stop It, London: Faber & Faber, 2022

[20] J. P. Crowley, 'Expressive Writing to Cope with Hate Speech: Assessing Psychobiological Stress Recovery and Forgiveness Promotion for Lesbian, Gay, Bisexual, or Queer Victims of Hate Speech', *Human Communication Research 40 (2013), 238–61.*

[21] UK Safer Internet Centre, 'Creating a Better Internet for All: Young People's Experiences of Online Empowerment and Online Hate', Lon- don: UK Safer Internet Centre, 2016.

[22] T. Keipi et al., 'Exposure to Online Hate Material and Subjective Well- being: A Comparative Study of American and Finnish Youth', *Online Information Review 42 (2018), 2–15.*

RESPONSE

HateLab and Mishcon de Reya published an **extensive overview** of the current legal and operational responses to online hate speech. These responses are limited in their effectiveness. For example, imposing large fines on platforms that refuse to remove hate speech will only work in a limited number of cases. The hate speech would likely have to be clearly criminal in nature for a take-down notice to be issued, leaving a wide array of grossly offensive content untouched (note that incidents motivated by gender-based hostility are not considered hate crimes in the UK). Imposing a 24-hour time frame on social media companies for the removal of illegal hate speech, as has been suggested by some, also means the damage is likely already done to the victim and the wider community. Even improvements in policing and prosecutions are unlikely to deter the most hardened haters, or those who post in the heat of the moment.

This issue is not a technical or legal one, but a social one. Fans must make a stand against hate.

COUNTER-SPEECH

We have the ability to coordinate in powerful ways to stop online hate. In the face of hate and abuse, counter-speech that reinforces community standards can change online behaviour, and perhaps the minds of those behind the screens. Counter-speech is any direct or general response to hateful or abusive speech which seeks to undermine it. Every social media user can favourably influence discourse through counter-speech by having a positive effect on the speaker, convincing them to stop propagating hate speech or by having an impact on the audience – either by communicating norms that make hate speech socially unacceptable or by 'inoculating' the audience against the speech so they are less easily influenced by it.

Combating hate speech with counter-speech has some advantages over law enforcement and platform sanctions: i) it can be rapid, ii) it can be adaptable to the situation; and iii) it can be employed by any internet user. Counter-speakers are often first at the online scene who witness the hate bubbling up. They are the 'online first-responders'.

Research provides evidence that counter-speech directed at the meso-level, in other words, not just at individuals or the entire hate network but at clusters of online haters (such as Facebook groups, communities and pages), can be effective. This effectiveness is bolstered when counter-speech is used across all platforms, and not just a few. By targeting only 10 per cent of hate network clusters, counter-speakers can destabilise the whole online hate network.[23]

HateLab is testing the effectiveness of different types of counter-speech sent to those posting hate on Twitter. Six forms of counter-speech are being considered:

- Attribution of prejudice to induce shame e.g "Shame on you for spreading sexist tropes like that! Imagine if someone said that about your daughter."

- Claims making and appeals to reason e.g. "This has nothing to do with gender. Take a look at these statistics…"

- Request for information and evidence
  e.g. "How does this have anything to do with gender?? Do you have any proof?"

- Jokes/comedy and reintegrative shaming
  e.g. Oh no, this woman sounds REALLY scary. I'm going to join an online incel group immediately - you guys. LOL!

- Mimicry and sarcasm highlighting issues with logic and consistency
  e.g. Hate speech: "I'm officially scared of butch lesbians. #NotHomophobic #JustScared"
  Mimicry: "I'm officially scared of bigoted men. #StereotypingMuch? #JustStupid"

- Reductio Ad Absurdum (logical argumentative philosophy that deploys an argument pushed to its logical absurd extremes to identify its inherent problem)
  e.g. "I guess what you're saying is you would feel more comfortable if women didn't exist? Fascinating. Do you want the human race to go extinct?"

[23] N. F. Johnson et al., 'Hidden Resilience and Adaptive Dynamics of the Global Online Hate Ecology', Nature 573, 261–5, 2019.

“FANS MUST MAKE A STAND AGAINST HATE”

Initial results show that counter-speech is effective in stemming the length of hateful social media threads when multiple unique counter-speech contributors engage with the hate speech poster.[24] However, not all counter-speech is productive, and evidence shows that individuals that publicly use insults against hate speech producers often inflame the situation, resulting in the production of further hate speech. When engaging in counter-speech, or advising others on its use, the following principles should be followed to reduce the likelihood of the further production of hate speech:

1. **Avoid using insulting or hateful speech**
2. **Make logical and consistent arguments**
3. **Request evidence if false or suspect claims are made**
4. **State that you will make a report to the platform, police or a third party if the hate speech continues and/or gets worse (e.g. becomes grossly offensive or includes threats)**
5. **Encourage others also to engage in counter-speech**
6. **If the account is likely a fake or a bot, contact the social media company and ask for it to be removed**

While likely to be effective in some instances, general counter-speech is unlikely to stem the production of hate in social media users that adopt extreme attitudes and worldviews. Those most susceptible to the stemming effects of counter-speech are those who use hate speech only occasionally (for example, only around 'trigger events'), and those that are not on a pathway to radicalisation. Counter-speech is unlikely to be effective on some bots or fake accounts, given their control is either fully or partially automated by computer code and their designed purpose is to spread division.

Those that care the most about football need to become hate incident first responders, setting and enforcing the standards of acceptable behaviour both online and in stadiums. While this may be the more difficult of the proposed solutions, it will surely be the most effective and long-lasting should we enact it.

[24] 'A study of cyber hate on Twitter with implications for social media governance strategies'. *Conference on Truth and Trust Online*, 2019 arxiv.org (Cornell University) | https://arxiv.org/abs/1908.11732

METHOD

HateLab facilitates access to all open-source online communications, including Twitter, Reddit, 4Chan and Telegram, for the monitoring and countering of online harms, including abuse, threats, identity-based-hate and divisive disinformation. For this report social media posts were sourced from Twitter, Reddit, and 4Chan between 2nd May and 1st August 2022. For Twitter, a query was created for all tweets @mentioning player account handles to ensure that posts directly targeting the women's squad were captured. Twitter is a reliable platform for the examination of communications sent to players as all the England women's national team have established profiles for public engagement purposes. As players do not have public-facing accounts on Reddit and 4Chan, queries were set up to obtain posts which mentioned their full names.

We used bespoke machine learning classifiers within the HateLab Dashboard to classify misogynistic (F1-score .81) and homophobic (F1-score .80) content. Our classifiers are trained on gold-standard human-annotated data. To build the training datasets, four human coders independently evaluated posts to determine if they were offensive or grossly offensive based on their misogynistic or homophobic content. Only posts that achieved 75 per cent agreement (3 out of 4 coders) were included in the hate class of the training datasets. Content flagged as hateful by our trained algorithms was subject to independent manual inspection by two hate speech experts to reduce false positives. To identify posts containing hopeful (positive) content we employed Google's Perspective API and used a threshold of >.85 on Valence Aware Dictionary for sEntiment Reasoning (VADER) scores.

We collected post metadata to identify levels of engagement with hateful content. Metadata from accounts found posting hateful content were collected to perform additional analysis on user sex and location, account creation date, and the number of followers and followees.

In accordance with Cardiff University's Ethics Committee Standards, we have avoided directly quoting hate speech posts to preserve the anonymity of social media users.[25] Instead, we present content in aggregate form via the visualisation of post content in tables, charts, word clouds, and emoji clouds.

THE BENEFITS OF HATELAB'S APPROACH

- **Blended AI and HI:** We partner with experts in civil society, government and law enforcement organisations to source online harms training data allowing us to continually update our AI to the highest standard
- **Minimal error:** We do not rely on simple keywords to identify online harms. Our natural language processing machine learning techniques are top performing, award-winning and verified in international peer-review open science journals (e.g. IEEE, ACM, WWW)
- **Deep knowledge:** We are world-class experts in hate speech, hate crime and cyberisk, and our founders are in the top 3 most cited in their fields
- **Rapidly adaptive:** Our AI + HI approach ensures we remain up-to-date with changes in online behaviours that can avoid detection in fully automated systems
- **Predictive:** Network and information propagation statistical modelling allow us to project the spread and survival of online harms, enabling enhanced threat assessment and mitigation

[25] Twitter Terms of Service forbid the anonymisation of tweet content (screen-name must always accompany tweet content), meaning that ethically, informed consent should be sought from each tweeter to quote their post in research outputs. However, this is impractical given the number of posts generated and the difficulty in establishing contact (a direct private message can only be sent on Twitter if both parties follow each other). Therefore, it is not ethical to directly quote tweets that identify *individuals* without prior consent.

HateLab is a nonprofit with an ambitious civic mission to democratise the latest AI and data science capabilities amongst civil society organisations so that they can reliably monitor and counter online hate speech, abuse, threats and divisive disinformation.

Arron Cullen is Head Analyst at HateLab. He holds an MSc in Terrorism, International Crime and Global Security and a PhD in Criminology. His research focusses on the nature of online hate speech, responses to online harms, and police use of social media. His latest published work appears in the *British Journal of Criminology*.

Matthew Williams is Professor of Criminology at Cardiff University and is widely regarded as one of the world's foremost experts in hate crime and online hate speech. He advises and has conducted research for the UK Home Office, the Ministry of Justice, the Foreign, Commonwealth & Development Office, the US Department of Justice, Google, Deutsche Telekom, EE and BT among others. Williams is also founder and director of HateLab, and he has conducted the largest dedicated study of hate victimisation in the UK. His research has appeared in documentaries for BBC One (*Panorama*, *Crimewatch*), BBC Two, BBC Radio 4 (*File on 4*), ITV (*Exposure*), Channel 4, CBS, Amazon Studios, and Netflix, and in major publications including the *Guardian*, the *Observer*, the *Independent*, the *Times*, the *Herald*, the *Los Angeles Times*, *Scientific American* and *New Scientist*. In 2021 he published the popular science book *The Science of Hate: How prejudice becomes hate and what we can do to stop it*, with Faber and Faber. @MattLWilliams

CARDIFF UNIVERSITY PRIFYSGOL CAERDYDD
UKRI Economic and Social Research Council