The Digital Misogynoir Report:
Ending the dehumanising of Black women on social media
Glitch is an award winning UK charity (1187714) committed to ending online abuse and making the internet a safer place for everyone, particularly women and unapologetically centring Black women.

Recognised internationally for mainstreaming the conversation on online abuse and digital citizenship, we are transforming how tech companies build and scale social media platforms so they have women's safety and joy in mind.

From delivering workshops, training and resources to communities all around the world, to trailblazing research reports, powerful change-making campaigns and working with global brands, we’re here to help make digital citizens of everyone.

Specific shoutout to Gabriela de Oliveira and Dr Julia Slupska for being co-authors of this report

In collaboration with

Textgain specialises in the development of AI that automatically detects and monitors harmful societal trends and tensions, such as hate speech and disinformation.

SUGGESTED CITATION:

Glitch, UK (2023) “The Digital Misogynoir Report: Ending the dehumanising of Black women on social media” www.glitchcharity.co.uk/research

This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (not inclusive of images used), which means you are free to:

**Share**

Copy and redistribute the material in any medium. Licensor cannot revoke these freedoms as long as you follow the licence terms, which are:

**Attribution**

You must give appropriate credit, provide a link to the licence, and indicate if changes were made. You may do so in any non-commercial reasonable manner, but not in any way that suggests the licensor endorses you or your use.

**ShareAlike**

If you remix, transform, or build upon the material, you must distribute your contributions under the same licence as the original.

**No additional restrictions**

You may not apply legal terms or technological measures that legally restrict others from doing anything the licence permits.

To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/4.0/ or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.
The cover image is an AI generated piece of artwork, depicting an image of a Black woman slowly disintegrating into pixels. It’s an artistic depiction of the idea of a Black woman losing her humanity to the online world. Whilst this AI image is in the style of hyperreal art, she is not real, and her likeness to any real living person is not intentional. This AI generated image was made for us by Visualist. It was made for us so that we did not need to use the stock image of a real Black woman, without being able to directly ask her permission to be the face of a topic on the dehumanisation of Black women online. That being said, there is a broader conversation to be had about AI art, particularly as it pertains to the ethical use of people’s images online.
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Vision &amp; Foreword</td>
<td>5</td>
</tr>
<tr>
<td>Content Warning</td>
<td>8</td>
</tr>
<tr>
<td>Executive Summary</td>
<td>10</td>
</tr>
<tr>
<td>Introduction &amp; Background</td>
<td>21</td>
</tr>
<tr>
<td>Methods</td>
<td>24</td>
</tr>
<tr>
<td>Findings</td>
<td>29</td>
</tr>
<tr>
<td>Calls to Action</td>
<td>48</td>
</tr>
<tr>
<td>Conclusion</td>
<td>57</td>
</tr>
<tr>
<td>References &amp; Appendices</td>
<td>62</td>
</tr>
<tr>
<td>Appendix 2: What is a toxic post?</td>
<td>64</td>
</tr>
</tbody>
</table>
Our vision is one where systemic change *actually* happens.

To awaken a generation of digital citizens equipped to create and demand for safe online spaces for all. Where tech companies not only understand their responsibility, but take an active stand in preventing – not just reacting to – online abuse.
FOREWORD

FOUNDER & CEO

Glitch as a campaign and now charity was born out of my personal experiences of misogynoir.

Back then, conversations about online abuse and violence towards women were focused solely on gender. Feeling the conflicting combination of hypervisibility as a Black woman while also being made to feel invisible by tech companies and law enforcement: the term misogynoir so perfectly sums up my horrific experience in 2017. A powerful clip of me confronting racists offline went viral within weeks and was shared across social media platforms as inspiration for defiance and allyship.

Not only does the term aptly describe the types of abusive comments that I printed and shared with the police, it also captures how dehumanising tactics rob us of Black joy, expression and resistance. The viral clip soon became an online game that racists could play with; the aftermath overshadowed a collective moment of liberation and then that became the story.

Black culture, particularly online, is beautiful, joyful, wise and certainly funny. It also aids the resistance of white supremacy, from dance trends to brave acts of mobilisation around issues of social justice. These self-organised safe havens for communities that are marginalised should be protected and supported to continue flourishing, rather than be met with hostility, attacks, and the hijacking of spaces and trends. Engagement on these platforms shouldn’t come at personal risk. These companies’ rapid expansion shouldn’t come at the expense of Black women’s safety, nor should we be required to give up so much of our personal data (we’ll need a whole other report for this).
This report illuminates the ways misogynoir shows up in online spaces; the way it spreads and intersects with other forms of white supremacy; and, most disappointingly, how it is still missed in content moderation by tech platforms. This report has been a long time coming; from influencing and supporting Amnesty International’s Toxic Twitter Report 2018, our own internal research on the abusive experiences that Black women Members of Parliament endured during the 2019 General Election in the UK and, finally, our Ripple Effect Report that looked at intersectional types of abuse. Now we have developed a dataset and clear policy recommendations for tech companies, governments, civil society, academia and the general public so that Digital Misogynoir is believed and is addressed. This report offers a roadmap for dismantling Digital Misogynoir and improving the lives of Black women, whilst acting as a stepping stone in driving transformative research so that we can provide better online frameworks for joy-centered safety, instead of merely surviving violence.

There has been a growing tolerance of dehumanising language towards Black communities, and at the very least an underestimation of the power that dehumanisation can lead to, both online and offline. Despite the glimpse of hope during Black Lives Matter 2020 this report serves to end the complacency and the ignorance of misogynoir, by illuminating an insidious form of prejudice that plagues our online spaces and which demands our immediate attention. Illumination chips at erasure, at gaslighting and validates the experiences of harm. Some of these harms are obvious (like death threats and slurs), but others are more indirect, like toxic tropes about Black women and femmes, or conspiracy theories which remain on platforms and at times go viral.

While there’s such a long way to go for AI as a tool to capture the complexities and nuances of misogynoir, the diversity within “Black women” and Black joy online, the spirit of this report is to present the reader with critical data of digital misogynoir in a trauma-informed and intersectional way. By doing so, my hope is that women, and particularly Black women, are not questioned when sharing their lived experiences, but just believed.

To Black women online, thank you for your resistance and community. I hope this report serves as an evidence tool so that you do not have to relive trauma, or be at risk of re-traumatising yourself, in order to be heard.

To non-Black readers, I hope, reading this report, you will agree that there can be no more excuses. No more unintended consequences when it comes to building and scaling social media platforms; no more research that doesn’t centre the most marginalised. I hope this evidence will reignite compassion and action, to be active digital citizens and help us mobilise to hold tech institutions accountable.

Together, let us challenge the existence of misogynoir, and build a path towards a future where Black women’s voices are heard, respected, and valued—both online and offline.

Seyi Akiwowo, Founder and CEO
Applying trauma-informed practices to this work has meant staying grounded in the experiences of harm that are represented throughout this research. This includes carefully considering what is shared and how we present it. In light of this, we have made the decision not to include the most explicit and violent abuse we analysed. Where toxic and harmful content is shared, we have blurred out the most harmful words.

This report does still include some explicit and harmful language, and an analysis of the following themes that may be distressing:

- Violent misogynoir and misogyny, including death threats and sexual assault
- Racism
- Homophobia
- Transphobia
- Antisemitism
- Islamophobia
- Violence

If you are impacted by any of the content of this research, you may find the following resources helpful:

**Glitch:** [Digital Self Care Resources](#)
**Coalition Against Online Violence:** [Online Violence Response Hub](#)
**PEN America:** [Online Harassment Field Manual](#)
**Mind:** [Racism and mental health](#)
**Wikimedia:** [Online harassment resource guide](#)
‘Black women have been shouting from the rooftops about this issue for not just years but decades. The writing has been on the wall for some time, but some have turned a blind eye to it. The truth is, online abuse has only become an issue since it began to affect the Global North. Specifically white middle-class women, and some men. But tech, governments, philanthropists, the media, the police and civil society all need to listen to and start centring Black women’s experiences online when shaping policies and legislation.’

-Seyi Akiwowo, How to Stay Safe Online, 2022

The developers of these apps are well rested billionaires who participate in the leisure life. We arguing with strangers on here and the rent still due.
EXECUTIVE SUMMARY

Black women* have been raising the alarm on digital misogynoir for years. Misogynoir, or the “particular venom directed at Black women”, is a term coined by Dr. Moya Bailey – with both Bailey and Trudy (aka @thetrudz) having significant roles in the creation and proliferation of the term2 – to acknowledge the ways Black women are uniquely discriminated against because of their gender and race.3 Digital misogynoir is the continued, unchecked, and often violent dehumanisation of Black women on social media, as well as through other forms such as algorithmic discrimination. Digital misogynoir is particularly dangerous because of its ability to incite offline violence. For example, after spending time on far-right social platforms, white supremacist Dylann Roof went on to murder nine Black church members, seven of whom were women, while they were at bible study.4 In the UK, misogynoir has recently been prominent in the sustained and targeted harassment of Meghan Markle in the tabloid press and online.5 Yet, the majority of online safety research and policy across civil society, government and tech companies continues to ignore the compounding racialised and gendered nature of online harm and abuse, with incredibly sparse work focussing on digital misogynoir. As a result, this report draws on a long tradition of Black feminist scholarship and activism to provide a statistical analysis of digital misogynoir. In collaboration with Textgain, a text analytics company that tracks hate speech, we analysed a large data set of posts from five major social media platforms: Facebook, Instagram, Twitter, Gab and 4chan. Building on our previous research in partnership with the End Violence Against Women Coalition6, we add to the existing evidence of the intersectional impacts of online abuse in a study that is the first to examine digital misogynoir across multiple online platforms. Through this work, we unpack what digital misogynoir looks like by providing further evidence for how Black women are disproportionately affected by online toxicity and abuse.

METHODS

We worked with a company named Textgain to collect and analyse a data set of social media posts about women. Textgain collected a total of 957,579 text-based messages posted between July 2022 and January 2023 from five different social media platforms: Twitter, Facebook, Instagram, 4chan and Gab, primarily using the European Observatory of Online Hate (EOOH) dashboard.7 Messages were collected using keywords related to neutral terms and slurs related to Black women, white women, and women in general. Textgain used a tool which detects toxicity using a lexicon-based approach, i.e. a dictionary of toxic terms. This means words and phrases are labelled by human annotators with a toxicity category (‘racism’, ‘sexism’, ‘religious-hate’, etc.) and a scale of toxicity from 0 to 1. To determine toxicity, an algorithm uses these tags from human annotators to assign scores to messages (see Appendix 2: What is a toxic post?). This allowed us to detect and compare toxic content across various categories and platforms. We also used a technique called topic modelling to identify underlying themes and topics in the data set. We supplemented this mainly quantitative analysis with a separate qualitative analysis on social conversation data conducted by Methods + Mastery, a social data agency.8 They analysed Twitter and TikTok data using both automated social listening tools and a manual ethnographic approach.

* See definition in Key Terms section
Our findings provide further evidence that digital misogynoir is toxic, dehumanising, and unchecked by major social media platforms, amplifying calls for greater tech accountability.

These companies can do so much more to upgrade their outdated content moderating systems to limit and prevent hate, but have made an active choice to scale up quickly without putting effective safety measures in place, putting profits over people.

1 Misogynoir, and misogyny more generally, were prevalent across all five social media platforms studied:

- Out of almost 1 million posts about women collected, 20% (1/5) were highly toxic. This amounts to over 1000 highly toxic posts per day.
- Posts about women are significantly more toxic than the average social media post.
- We, as external researchers, have fewer resources, tools, and data at our disposal than major companies on their own systems – why are we still finding this much abuse going unaddressed?
Hateful tropes continue to be used to silence and harm Black women, but social media can also be used by and for Black women to challenge abuse and build community

- **a.** We found over 9000 more highly toxic posts about Black women than white women in a data set which highlights the way that Black women are more likely to be racialised, i.e. referred to in reference to their race or ethnicity.

- **b.** Dehumanising language and stereotypes long critiqued within Black feminist scholarship - such as ‘the angry Black woman’ (‘Sapphire’), fetishisation (‘Jezebel’), and fatphobia (‘Mammy’) – are rife in digital spaces.

- **c.** The most prevalent misogynoir trope that was found was that of the ‘angry (fat) Black woman’ – an extremely harmful (and fatphobic) trope that often leads to both dismissal and tone policing of Black women’s voices in public discussions.

- **d.** Despite this toxic misogynoir, we also see the power and joy of Black online communities in how social media is used to challenge abuse and celebrate Black women which is shown in the relatively high rates of positive content about Black women.
Misogynoir underpins hateful narratives like white supremacy, antisemitism, and great replacement theory*:

Hateful rhetoric and jargon is trickling from the alternative platforms (Gab, 4chan) to the mainstream ones (Twitter, Instagram, and Facebook).

a. On mainstream social media, we mostly find stereotyping, body-shaming and fetishising. In contrast, on alternative platforms, we find more white supremacist and antisemitic themes.

b. Hateful jargon from alternative platforms - like ‘gorillion’ and ‘globohomo’, and conspiracy theories like ‘the great replacement’ - are steadily trickling into the mainstream.

c. Even abuse aimed at white women is often based on demeaning other races e.g. racist vitriol against mixed race couples in which white women are seen as ‘betraying the white race’

* See “Great replacement theory” in Key Terms.
TECH COMPANIES

1. **Include a clear definition of and policies on misogyny, misogynoir, racism and white supremacy** in their work on platform design, safety and content moderation and specific online gender-based violence prevention.

2. **Release transparency reports with disaggregated data available for independent research and analysis** on content moderation, account takedowns, wider enforcement and appeals. These are to include race, gender, and other identity characteristics of abusive posts, flagged posts, and content takedown. They should also include more detail about which posts are flagged by automated moderation, active bystanders and the targeted users themselves.

3. **Commit to removing barriers to dataset access for independent research** with consistency and transparency. Tech companies need to stop clamping down on tools like Crowdtangle and APIs which allow researchers to collect social media data. We also need consistency in dataset access across platforms. We need access to image and video content in order to analyse and moderate deepfake generated content.

4. **Collaborate with experts who can critique and improve online safety policies and abuse detection methods from a Black feminist standpoint.** There is a need for transparency, human input, and expertise from a variety of lived experiences in AI annotation, implementation and interpretation.

5. **All platforms commit to a well resourced Trust and Safety Council that has a key priority pillar on gender-based violence** that cuts across themes of age, verification, privacy as well as harmful content.
Implement a public health approach to online gender-based violence

a. Invest in civil society and academic research to better understand the links between forms of offline and online hate such as misogyny, misogynoir, incel rhetoric, and white supremacy

b. Define white supremacy and incel ideologies as forms of extremism

c. Focus on better prevention and education rather than carceral responses to extremism, including an awareness raising public health campaign around misogyny, racism and misogynoir to educate people across society, not just those still in formal education

Hold tech companies accountable on misogynoir, including:

a. Through up-to-date regulation and legislation on gender-based violence including strong provisions on content moderation and transparency reports

b. Comprehensively include smaller platforms in online safety legislation in relation to online harms, with a particular focus on platforms which act as a conduit to white supremacy

c. Implement a ‘tech tax’ following a polluter pays principle. 10% of this tax should be ring-fenced to support the work of civil society organisations doing vital work in this area

Pass and effectively implement future-proofed intersectional legislation on online safety: integrate misogynoir and online abuse into broader workstreams work around gender-based violence, as well as countering radicalisation. This legislation should name online gender-based violence and/or violence against women and girls explicitly and take a gendered, intersectional approach to preventing and mitigating harms from a systems-based approach that champions safety-by-design

a. Invest in programmes that focus on the recruitment and retention of women in science, technology, engineering and maths, particularly Black and minoritised women, to address the gender and diversity imbalance within the UK tech sector

* Including, but not limited to, the European Union (EU), United Nations (UN), and the Organisation for Economic Co-operation and Development (OECD).
Work together across sectors: because abuse is not ‘just tech rights’, ‘just misogyny’ or ‘just racism’

a. Digital rights work must include Black feminist analysis

b. Women’s rights work must include digital analysis, and an intersectional analysis of all Black women across the gender spectrum (cis and trans)

Focus on research in other languages – English content moderation is by far the most invested in worldwide / platform wide in text, images, and video – yet we still see huge gaps, so what are the implications for other languages?

Adopt intersectionality as best practice in online safety research: focus on marginalised groups who are often lost/excluded in one-size-fits-all safety work

a. In addition to influencers, politicians and other hyper-visible people in the public eye, we need to focus on everyday minoritised people who do not have the social capital to challenge their abuse

b. Prioritise Black trans women, and understand that some Black nonbinary, agender and gender-variant folks, as well as Black trans men are marginalised in the digital space because they may be perceived within a gender binary and therefore subjected to misogynoiristic harm

Fund research for, by and about Black women: fund Black women to do research degrees and fund those already in research

a. This kind of research is still too often dismissed as ‘too niche’ – which is itself a form of institutionalised misogynoir, given the extreme levels of risk and abuse Black women face.
The Digital Misogynoir Report

The men who are radicalised into ideologies like incel hate (which harms both them and others around them). We also highlight a need for civil society, communities, and research to be part of the design and platform moderation processes. Digital platforms have known they have a problem with online abuse, misogyny, racism, and radicalisation into conspiracy theories for years, not least because Black women have been raising the alarm on it since the 1990s. It is a choice to delay, ignore, or deny the problem. It is a choice not to take more concerted action against digital misogynoir and other forms of online hate.

This is why we have a set of recommendations in our Calls to Action for tech companies to make this mandatory and no longer something they can opt-in. We also lay out tangible steps that governments and intergovernmental organisations; research and civil society organisations; online communities and digital citizens can take to hold tech companies accountable and contribute to the dismantling of misogynoir. We need action from the top down (i.e. at an institutional level) and from the bottom up (i.e. grass roots).

**CALL TO ACTION**

**COMMUNITIES/DIGITAL CITIZENS**

1. Understand the harmful effects of misogynoir, including the racist and sexist tropes that underpin it

2. Practise good digital citizenship and digital self care: challenge misogynoir. Vocally support Black women online. Take care of yourself and others by setting digital boundaries

3. Follow and listen to Black women: build community with different people and amplify Black women across intersecting identities

4. Demand better from tech companies through collective action – sign up to Glitch’s newsletter to join our upcoming campaigns

Without safety for Black women, there is no safety for all.

This research makes clear that tech companies are not doing enough to include safety-by-design in all their products to prevent and reduce harm. We call on tech companies to have specific policies on misogynoir, outlining how they will prevent and reduce the harm caused by misogynoiristic content on their platforms. More needs to be done to safeguard against misogynoir, including increasing the effectiveness of content moderation – using both machine learning and human moderation in a holistic way. Tech companies should be investing in resourcing preventative measures related to misogynoir reduction. Ignoring abuse, toxicity, and harm against Black women means current efforts by tech companies and governments to challenge radicalisation and make online platforms safer for everyone are missing a key part of online hate. Aside from highlighting the gap in protections for Black women in online spaces, this report further evidences the critical need for online safety efforts to see Black women’s wellbeing as fundamental to broader collective change. This is both for the benefit of Black women, as they are disproportionately abused online, and for other communities such as women in general, Jewish people – some of whom are Black women – and the men who are radicalised into ideologies like incel hate (which harms both them and others around them).
#MISOGYNOIR
A term that Dr. Moya Bailey and Trudy (aka @thetrudz) having significant roles in the creation and proliferation of, which is used to acknowledge the ways Black women are uniquely discriminated against because of their gender and race, resulting in ‘particular venom directed at Black women’. Misogynoir specifically puts anti-Blackness at the core of its analysis of gendered racism, giving us space to explore the experiences of Black women*. Misogynoir can be expressed through erasure of Black women’s experiences, hypersexualisation and objectification of their bodies, dehumanisation, and silencing of their voices and perspectives.

*Note: In keeping with Dr. Bailey’s work, we understand ‘Black women’ as a diverse range of people from different countries, ethnic groups, religions, schools of thoughts and ways of life, and both trans and cisgender women. Dr. Bailey states that “Black women’ is often assumed to mean straight and cis, with queer and trans Black women identified explicitly because of this normative assumption… ‘Black women’ is not inclusive of nonbinary, agender, and gender-variant Black folks whose experiences of misogyny are intimately connected with a misgendering of them” As Dr. Bailey notes, “‘women and femmes’ doesn’t quite capture all the targets of misogynoir. There are masculine-of-center, agender, and nonbinary people who experience the deleterious effects of misogynoir and who may not identify as women or femmes.”

Therefore we repeat Bailey’s call:

“I challenge you, dear reader, as you read this text, to think of Black women first when you see the word ‘woman’, to think of queer and trans women first when you read the term ‘Black women’.”
#INTERSECTIONALITY
“Intersectionality” was introduced by Ann Julia Cooper as far back as 1892 and was popularised by American scholar Professor Kimberlé Crenshaw to refer to the way different aspects of a person’s identity intersect. According to Lisa Bowleg, “intersectionality is a theoretical framework for understanding how multiple social identities such as race, gender, sexual orientation, and disability intersect at the micro level of individual experience to reflect interlocking systems of privilege and oppression (i.e. racism, sexism, heterosexism, classism) at the macro social-structural level.” In addition to speaking about intersectionality, Glitch also refers to the concept of “multiple identities,” acknowledging that “women who face multiple and intersecting forms of discrimination offline because of their different identities (i.e. race, ethnicity, religion, sexual orientation, gender identity, disability, etc.) are also likely to be targeted with discrimination that targets their multiple and intersecting identities online.”

#ANTI-BLACKNESS
In ‘Me and White Supremacy’, Layla F Saad uses the Merriam-Webster definition of ‘anti-Black’: “opposed to or hostile towards Black people” and the Movement for Black Lives defines anti-Black racism as a “term used to specifically describe the unique discrimination, violence and harms imposed on and impacting Black people specifically.” Contemporary violence and discrimination is what Saidiya Hartman identifies as an afterlife of slavery – anti-Blackness normalised over five centuries, resulting in premature death and skewed life chances. In ‘America, Goddam’, Treva Lindsey shows how Black women and girls face unique forms of anti-Blackness interconnected with misogyny and capitalism.

#WHITE SUPREMACY
White supremacy is an ideology which maintains that white people are superior to other groups; this ideology is expressed through rhetoric, policies, and practices that reproduce the dominance of white people and the oppression of people of colour, especially Black people. While white supremacy is most often associated with overt forms such as lynching, hate crimes, police brutality, and overtly racist jokes, it also appears in a variety of more subtle but deeply pernicious ways: hiring discrimination, claims of “reverse racism” and forms of minimisation through irony and humour such as blackface (“it’s just a joke!”). Understood this way, white supremacy – particularly in countries like the US and UK – is “a central organising principle of social life rather than merely an isolated social movement.” The racial resentment underlying white supremacy is linked to broader social and economic factors, as W.E.B. Du Bois argues, racial “dislikes, jealousies, hatreds” persist and are encouraged because the very wealthy benefit when white working class people compete with Black working class people, leading to division and murder rather than organised resistance against exploitation.
#GREAT REPLACEMENT THEORY
Great replacement theory, ‘white genocide’, or simply ‘replacement theory’, is a white nationalist far-right conspiracy theory originating in France that maintains that the ethnic French and white European populations at large are being demographically and culturally replaced by non-white peoples – especially from Muslim-majority countries – through mass migration, demographic growth, and a drop in the birth rate of white Europeans. The theory has antisemitic roots as well, often pointing to the complicity or cooperation of ‘replacist’ elites (often a dog whistle for antisemitic conspiracy theories). While similar themes have been present in various far-right theories since the late 19th century, the particular term was popularised by Renaud Camus in his 2011 book *Le Grand Remplacement*. This conspiracy theory has been linked to several mass-shootings aimed at minoritised and immigrant communities. 

#TOXICITY
Online toxicity encompasses rude, aggressive, demeaning and degrading attitudes and behaviour on online platforms (in this report, social media platforms). This includes hate speech targeting protected characteristics such as race, gender, or religion, as well as abusive language such as death threats or threats of sexual assault. Following a trauma-informed approach, we have not included screenshots of toxic speech in the main body of the report. However, you can see examples of toxic posts in Appendix 2: What is a toxic post?

#DEHUMANISATION
Dehumanisation is the denial of full humanity in marginalised groups: in other words, treating others as ‘less than human.’
This can occur through describing humans as animals (such as monkeys, rats, or vermin) or comparing them to objects (commonly applied in the sexual objectification of women). Dehumanisation as a rhetorical strategy has been used to incite genocide and justify war, slavery, and the denial of suffrage and other basic rights. In particular, dehumanisation of Black people was and is central to slavery, colonisation, and ongoing forms of racism.
INTRODUCTION & BACKGROUND

Black women face persistent and disproportionate abuse online. A report by Amnesty International found that Black women are eighty four percent more likely than white women to be on the receiving end of online harassment.29 This reality is missed in most systemic responses to online abuse, which usually only mention misogyny if they discuss identity factors at all. Studies of online abuse that focus on misogyny or ‘just gender’ do not take into account the unique and specific ways Black women experience harassment, which is similar in some ways, yet so different.30 Digital misogynoir is the continued, unchecked, and often violent dehumanisation of Black women on social media, as well as through other forms such as algorithmic discrimination. The fact that the tech safety sector has failed to address this problem is an issue of ‘silenced knowing’, a phrase used to denote how knowledge held by Black women is often wilfully ignored and silenced.31

Dr. Moya Bailey outlines how digital misogynoir draws on racist and sexist tropes about Black women, such as:

#Sapphire: a mean, emasculating woman, ordering Black men to do her bidding (also seen in the stereotype of the ‘angry Black woman’ or the ‘sassy Black friend’)

#Jezebel: a promiscuous and highly sexualised woman

#Mammy: a sweet, caring, stupid, fat (i.e., undesirable) caregiver

Criminal and/or “#welfare queen” a stereotype of Black women committing welfare fraud constructed in the late 1970s by US media outlets and politician officials to justify federal welfare cuts32

These harmful tropes combine misogyny and racism with other forms of oppression, such as fatphobia and fetishization. Naming and recognising these tropes allows us to challenge the systems that marginalise Black women online and offline. Black feminist theorists like Dr. Patricia Hill Collins, bell hooks, and Dr. Melissa Harris-Perry have shown how these tropes or ‘controlling images’ attempt to delimit the potential ways of being for Black women in the world.32 These images serve the hegemony of “white supremacist capitalist patriarchy” by dictating the way society views marginalised groups and how we view ourselves.33 As Moya Bailey points out, misogynoir is not only a matter of stereotyping in popular culture, but also underpins destructive policies like the criminalisation of Black women and children, welfare cuts, and the systemic failures to address Black women’s health in public health policy. Authors such as Dr. Ruha Benjamin34, Dr. Guilaine Kinouani35, and Dr. Annabel Sowemimo36 have also extensively documented the adverse impacts of racism and misogynoir on Black women’s physical and mental health.

However, Bailey also spotlights the myriad forms of Black women’s resistance to digital misogynoir. Black women, and particularly Black trans and queer Black women, have used the internet and social media to create and share innovative stories outside mainstream structures of media funding, to challenge oppression and discrimination in their communities, and to

“transform misogynoir by building powerful digital and real-world networks.”37
Bailey’s research follows a long tradition of intersectional investigations (see ‘Intersectionality’ in the Key Terms) into oppression both online and offline. Bailey’s work focuses on US visual and digital culture: we also need more work on misogynoir’s effects on the Black diaspora outside of the US, to better understand the complexities and nuances of Black experiences and build a stronger collective movement towards systemic change. Despite the fact that intersectionality has been debated since the 70s, most large scale quantitative studies of online harassment do not consider intersectional ways abuse manifests online.

For example, while the Pew Research Center’s 2021 ‘State of Online Harassment’ report notes the prevalence of gender-based and race-based abuse, it does not include overlapping impact. Similarly, a systematic review of evidence on online abuse by the Turing Institute does not include any mentions of intersectional impacts. In fact, we found it difficult to get specific funder support for this research, as Black women’s safety is considered by many funders to be too ‘niche’. We need further funding to expand research on the basis of this exploratory work.

This report, therefore, builds on a small but important evidence base on the intersectional nature of online abuse. In 2018, Amnesty International investigated harassment of women politicians and journalists in the UK and USA, finding that Black women were 84% more likely to be harassed with racist and sexist abuse and threats of violence. In 2020, we published a report in collaboration with the End Violence Against Women (EVAW) coalition, which showed that Black and minoritised women and non-binary people were more likely to experience online abuse during the Covid-19 pandemic. This kind of quantitative evidence is crucial to document the ongoing scale and prevalence of the problem, which persists despite years of attempts to counter hate speech and improve online moderation. However, we note that large scale quantitative evidence always carries limitations (explored further in the limitation section of our Methods), particularly in capturing the nuances and complex impacts of abuse, and therefore we wish to point interested readers to excellent qualitative work on digital misogynoir by Dr. Moya Bailey and others such as Dr. Kishonna Gray. In her report to the UN Forum on People of African Descent, Mutale Nkonde provides a thorough evidence base on algorithmic bias against Black women, documenting both online hate and how it can incite offline violence.

Research on misogynoir detection is in very early exploratory stages. Researchers at the Open University conducted a study on four cases of Black women in tech reporting experiences of misogynoir on Twitter. Kwarteng et al. (2021) collected reactions to these cases (both supportive and non-supportive), and categorised them within a model of misogynoir that highlights experiences of tone policing, white centring, racial gaslighting and defensiveness. In a subsequent
study, Kwarteng et al.\textsuperscript{46} analysed state-of-the-art
detection methods used by tech companies, such
as HateSonar and Perspective API, and found that
they are ineffective at detecting intersectional
hate such as misogynoir.\textsuperscript{46}
In particular, their qualitative examination of
false positives and false negatives revealed that
these systems were classifying many instances of
tweets containing references to racism, sexism,
and profane or aggressive language as hate
speech. This creates risks for self-advocacy or
the use of African-American English (AAE), which
may be inappropriately flagged as racist content.
These AI systems struggle to identify other subtle
types of hate and are insensitive to context,
which is a crucial component of misogynoir and
intersectional hate. Kwarteng et al.’s work\textsuperscript{46} also
highlights the importance of applying a Black
feminist theory to hate speech detection: “In a
Black feminist interpretation of racism, power
is an essential feature in determining what
is ultimately racism.”\textsuperscript{46} General approaches
which view all racial markers as hateful will flag
Black women’s counter-speech against “racism”.
This is a particular challenge for lexicon-based
approaches (like ours) as many of the same
words are used by both hate speech and counter-
speech. Lexicon-based approaches also struggle
to pick up the nuances of misogynoir, like tone
policing. The evaluation of models for detecting
online hate speech and toxicity is a relatively
new and rapidly growing field of research: a
lack of focus on intersectionality is one of many
challenges in the field.\textsuperscript{47}

We build on this work by providing the first large-
scale cross platform analysis of intersectional
forms of online abuse. By collecting posts related
to Black women, white women, and women in
general, we also move away from a focus on
harassment of individuals (including prominent
individuals) and move towards examining the
toxicity and harmful tropes that are proliferating
against women and particularly Black women
as whole groups of people. The dehumanisation
of Black women online is particularly troubling
as social media becomes included in the large
scale data sets used to train search algorithms
and generative AI, which means the risk that
racist, sexist, and otherwise abusive language
ends up embedded as ‘normal’ in AI models.\textsuperscript{48}
Dr. Safiya Noble\textsuperscript{49} illustrated the harmful effects
of this kind of algorithmic racism by showing
how Google searches for Black women are more
likely to result in harmful pornographic content
(reinforcing the fetishisation of Black women or
the ‘Jezebel’ stereotype outlined above). By raising
awareness of the prevalence of toxic tropes and
narratives now, we can stop digital misogynoir
before it proliferates and becomes even further
embedded in emerging digital technologies.\textsuperscript{50}

\textbf{PICKING UP NUANCES}

Those type of
girls are always
#sluts and give
it up easy.

This the only
night u can be
a #slut and not
get judged so i
need yall on ur
worst behavior

Lets get spooky
#Explorepage
#prty #party
#lit
METHODS

DATA COLLECTION

We worked with a company named Textgain to collect and analyse the data set. Textgain collected a total of 957,579 text-based messages posted between July 2022 and January 2023 from five different social media platforms: Twitter, Facebook, Instagram, 4chan and Gab, primarily using the European Observatory of Online Hate (EOOH) dashboard. We often refer to Facebook, Instagram and Twitter as ‘larger mainstream platforms’ given the size and reach of these platforms. We then refer to Gab and 4chan as ‘smaller alternative platforms’ to again reflect their size and reach. These platforms were chosen due to their prominence and because they are primarily text-based (unlike TikTok and YouTube). Although we included Instagram in this study (through the analysis of post captions), these results are limited as Instagram content is primarily visual.

Messages were collected using keywords related to neutral terms and slurs related to Black women, white women, and all women in general. We contributed words that were specifically related to misogynoir, including words and phrases from screenshots of abuse sent to our CEO (see Founder & CEO Forward), to keep this research grounded in lived experience of abuse.

Textgain then collected and analysed data across the ‘neutral’ and ‘toxic’ categories (see Figure 1). This allowed us to examine more subtle abuse like when Black women are referred to in toxic ways without the use of specific racialised slurs and whether this differs from how white women or all women in general are referred to. For all categories, the same number of keywords was used to collect data.

Textgain also collected a matching sample of completely random data from the same time period for each platform, in order to develop a ‘baseline rate’ of toxicity, against which we could measure the samples we collected using neutral and toxic terms for women.
**Keywords**
(for neutral messages)

- 'woman', 'women', 'girl', ...
- 'white woman', 'white girl', 'fair-skinned women', ...
- 'black woman', 'black girl', 'black skinned woman', ...

**Keywords**
(for toxic messages)

- 'b*tch', 'slut', 'roastie', ...
- 'n*gress', 'black b*tch', 'nappy head girl', ...
- 'white b*tch', 'mudshark', 'white trash wh*re', ...

**Important to note that some messages collected using neutral words are still toxic.**

**FIGURE 2: NUMBER OF MESSAGES COLLECTED PER PLATFORM**

<table>
<thead>
<tr>
<th>Platform</th>
<th>Neutral Messages</th>
<th>Toxic Messages</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Twitter</strong></td>
<td>333,903 (35%)</td>
<td>115,011 (12%)</td>
</tr>
<tr>
<td><strong>4chan</strong></td>
<td>217,919 (23%)</td>
<td>179,113 (19%)</td>
</tr>
<tr>
<td><strong>Facebook</strong></td>
<td>111,633 (12%)</td>
<td>23%</td>
</tr>
</tbody>
</table>
DETECTING TOXICITY

Online toxicity encompasses rude, aggressive, demeaning and degrading attitudes and behaviour on online platforms (in this report, social media platforms). Other terms – like online ‘abuse’ or ‘harms’ – can also be used, however we focused on ‘toxicity’ as it is broader, and because we are not collecting data specifically on people’s experiences of harm (just on toxicity rates in the messages we collected.) Toxicity in speech is necessarily subjective: different people will annotate toxic speech differently based on lived experience and differing worldviews. As there is no fully objective measure of toxicity, and there will always be a degree of disagreement among both people and algorithms as to how toxic certain messages are; we return to this in the limitations.

Textgain has developed a tool which detects toxicity using a lexicon-based approach, i.e. a dictionary of toxic terms. Textgain’s English lexicon has a dictionary of over 7,500 toxic words and word combinations. These words and phrases are labelled by multiple human annotators with one or more toxicity categories (‘racism’, ‘sexism’, ‘aggression’, etc.) and a toxicity scale from 0.0 to 1.0. To determine toxicity, an algorithm uses these tags from human annotators to assign scores to messages (see Appendix 2: What is a toxic post?). This allowed us to detect and compare toxic content in a large data set across various categories and platforms. Overall, the algorithm was fairly accurate at detecting hate speech, with an 83.6% accuracy rate of detecting posts marked as toxic by human coders. Where the algorithm made mistakes, it was more likely to underestimate the number of toxic posts than to overestimate them. Therefore, the statistics in the report should be understood as likely underestimating the levels of toxicity in our sample. Please see Appendix 3: Accuracy tests for a detailed discussion of evaluating this toxicity detection tool.

We also used a machine learning technique called topic modelling to identify underlying themes and topics in the data set. In this way, we were able to identify which topics on social media are repeatedly linked to Black women, white women and women in general in an automated way. For this, we used the Python libraries BERTopic and SentenceTransformers. Insights from the toxicity rates and topic modelling were then combined with qualitative readings of a subset of messages from each category.

We supplemented this quantitative analysis with qualitative analysis on social conversation data conducted by Methods + Mastery, a social data agency. They analysed Twitter and TikTok data using both automated social listening tools and a manual ethnographic approach. Lastly, earlier drafts of this report benefitted from the feedback from two excellent community reviewers, Mutale Nkonde (AI for the People) and Nani Jansen Reventlow (Systemic Justice).
LIMITATIONS AND AREAS FOR FURTHER RESEARCH

The data we collected has several gaps. As we discuss later in the findings, some platforms allow more data access, while others (like Facebook) have more private groups from which we could not access data. Therefore, the platform averages are not necessarily representative, but merely a reflection of data that we could collect. We did not investigate who saw each message, how many times they were viewed, how long abusive posts stayed online or what safety measures were applied by the platform. This data on content moderation is not available to researchers.

We did not collect data on the context surrounding each message, such as who it targetted or who created the misogynoir content we detected. It would be interesting, in further research, to better understand the backgrounds and identities of the creators of misogynoir content.

We focused primarily on toxic content about Black women, and compared it to toxic content about white women and women in general to allow us to unpack the relationship between misogynoir online and white supremacy. This is based on Black feminist practice, which seeks to centre a group who are often excluded in research and policy. This means we did not collect data on other groups of women of colour who also experience gendered and racialised abuse online, although our analysis does uncover overlaps between misogynoir and other forms of hate related to different groups of women of colour. This is an important area for future research.

The annotation process also has limitations. Automated tools like the one we used can include built-in bias and lack the sophistication to consistently distinguish between toxic content itself and ‘counter-speech’, i.e. content that discusses, responds to, or critiques toxic content. Black women have been disparately targeted by content moderation for speaking about racist and sexist incidents online and offline. The team of annotators had gender parity but were predominantly white, and therefore, may have missed some of the nuances of toxicity that annotators with lived experience of racism and sexism may have identified. In particular, there are some terms that are specific to the Black diaspora that may not be seen as toxic by non-Black people. This is unfortunately true across most research in this space, as most training data sets for research on hate speech detection depend heavily on crowd-sourced raters, who may lack knowledge of misogynoir. There are also benefits to this, as people from racialised groups may have found racist content more harmful or triggering to engage with. In contrast, our raters did receive specific training on misogynoir. Textgain’s lexicon-based approach is based solely on combining toxicity scores for words and, therefore, may miss many nuances in language (although we note the relatively high levels of accuracy in our accuracy tests). We have noted, in various parts of this research, where this may have impacted the data collected.
We used relatively simple methods compared to the enormous computational power and sophisticated tools for AI-enabled content moderation, which are available to major social media companies. Our content analysis focused only on text-based abuse and did not account for pictures or videos – many of which will have been highly toxic. Image-based forms of misogynoir – such as attaching a photo of a monkey to a prominent Black politician – would be missed in our analysis. This means we did not account for much of the content on more visual platforms, like Instagram, and that our methods would be poorly suited for predominantly video based platforms like YouTube or TikTok. Given private messages are also not analysed as part of this research, the data collected across all platforms is likely an underestimation of the relevant content across all categories and topics.

Detecting toxicity and misogynoir in visual media is an important developing area of research. Yet we know that content moderation is even more underdeveloped for video, images, and emoji, not to mention the metaverse and other forms of augmented reality, or the massive data sets that make up large language models. Companies should not push out these new technologies before there are tried and tested methods to prevent and redress harm on existing technologies.
FINDINGS

Our findings provide further evidence that digital misogynoir is toxic, dehumanising, and unchecked by major social media platforms, amplifying calls for greater tech accountability.

This section outlines 3 key findings:

1. Misogyny and especially misogynoir are prevalent across social media platforms

   We found over 200,000 highly toxic posts about women across all platforms studied, including:

   - 200,976 highly toxic posts
   - 154,373 Women in General
   - 27,874 Black women
   - 18,729 White women

2. First, using relatively simple methods, we found a large amount of highly toxic posts (over two hundred thousand) across all platforms. We show that social media companies are not successfully moderating content at the intersection of racist and sexist hate.

3. Second, we found that Black women receive a disproportionate amount of the most vitriolic kinds of toxicity. However, despite this toxicity, Black women are also celebrated by a joyful and supportive community of other Black women.

4. Third, we found that misogynoir underpins broader hateful narratives like antisemitism, white supremacy and great replacement theory.
Every category of data we collected, including those collected using neutral keywords, was also significantly more toxic on average than the random baseline data set, which was a sample of completely random posts from the same time period for each platform. This indicates that posts mentioning women are significantly more toxic than the average post on these platforms. This section goes further into findings between different platforms. Figure 3 shows mean toxicity by platform.

**FIGURE 3. DISTRIBUTION OF TOXICITY SCORES, BY PLATFORM**

Overall, of the platforms we analysed, 4chan is by far the most toxic platform in relation to misogynoir followed by Gab. As both these platforms have positioned themselves as ‘free speech’ and ‘anti-woke’ platforms, it is not surprising that they attract many users who have been banned from or moderated by mainstream platforms, including communities of neo-Nazis and white supremacists. However, Twitter is the most toxic larger mainstream platform overall (i.e. mean toxicity) Facebook has the highest proportion of the highest toxicity content of these companies. This means we were able to identify fewer toxic messages on Facebook than Twitter, but that these messages are more often extreme in nature.\(^*\)

The number of misogynoir posts found on the Meta platforms (Facebook and Instagram) is relatively small compared to their enormous user base. This is surprising as in previous research, Facebook has been reported as the platform where individuals have experienced the most abuse.\(^57\)

\(^*\)Noting that messages on Facebook are often much longer than messages on other platforms. This generally means more toxic words per message. This creates a higher likelihood that content will be rated with a higher toxicity. Equally, it raises a question around why these messages are not being moderated or prevented.
However, this is likely due to the difficulties of accessing data on these platforms, including the fact that we did not analyse visual data. For example, a common trolling technique to get around blocked terms and moderators is to post abusive images as comments. The tool we used for data access on Facebook (called 'Crowdtangle') only allows access to public groups to collect data. Most public groups seem to be monitored fairly well, as they have their own monitoring teams. This partly explains the low number of misogynoir messages found on Meta platforms. However, we could not access private groups, which is where truly toxic speech is more likely to proliferate. This points to a need for greater transparency from Meta, as routes for access, including ‘Crowdtangle’, have grown steadily more limited in recent years. Due to differences in data access, it is difficult to make cross-platform comparisons or to accurately rank these platforms in terms of their toxicity. We supplemented this quantitative analysis with qualitative analysis on social conversation data conducted by Methods + Mastery, a social data agency, which is flagged in call-out boxes throughout the findings.58

Methods + Mastery

SOCIAL PLATFORMS CAN MAKE VICTIMS FEEL POWERLESS.

The general consensus among victims is that #NO PLATFORM IS TRULY SAFE. People recounted their abusive experiences from Twitter, Reddit, TikTok and Instagram, as well as more niche platforms like Discord and even Roblox. Even things that appeared to be harmless – like viral TikTok challenges (i.e. “Tell me about your worst Tinder date”) – were triggering to some.

From being blasted with misogynistic comments to being subjected to months of online abuse, victims slammed social platforms for THEIR REFUSAL TO TAKE ACTION against abusers and for AN OVERALL LACK OF REGARD for people’s wellbeing online.

The trans community, in particular, expressed disappointment at the lack of support from Twitter and criticised the platform's lack of effort to protect the community.

- Victim On Twitter

I’ve seen many trans people report purposeful misgendering which contributes to our poor survival rates to be IMMEDIATELY DENIED BY TWITTER STAFF that nothing was going on with 0 in person review so apparently they don’t

- Bystander on Twitter

Automated replies which dismiss reports of abuse (“0 in person review”) can be particularly invalidating. When reports by victims or bystanders were taken seriously by platforms, reactions were a mix of shock & elation.
This means that social media companies are not successfully moderating content related to women—particularly content at the intersection of racist and sexist hate. We have less resources and tools and data at our disposal than these major companies. Thus, the fact that this study still detected such high quantities of abuse points to a significant lack of investment and expertise in preventing and moderating misogynoir. Black women have raised this problem for decades. Tech companies know about this. Why are these tech companies still allowing this content to persist, often in violation of their own policies? We return to this discussion in our section on recommendations.

2 Hateful tropes continue to be used to silence and harm Black women, but social media can also be used by and for Black women to challenge abuse and build community.
Figure 4 shows the messages collected using neutral and toxic keywords related to Black women, white women and women in general. Overall, there is a lot more content referring explicitly to Black women than to white women. This shows how Black women are more likely to be racialised, meaning they are more likely to be spoken about in relation to their perceived race or ethnicity, whereas white women are less likely to be referred to in terms of their race and more likely to be referred to simply as women.

Conversely, when people speak of just ‘women’ online, they often miss or ignore the experiences of Black women, implicitly treating white women as the default. Messages that are assigned a toxicity score above 0.8 are highly likely to be extremely offensive. Posts about Black women collected using toxic keywords include by far the highest proportion of the most toxic content. However, despite these high rates of toxicity detected in relation to Black women, we also saw many positive posts about Black women, particularly in the data set collected using neutral keywords. This section dives into differences in toxicity among different categories, with particular focus on the most toxic posts (ie. those with a toxicity score above 0.8). We start with the data set collected using slurs as keywords, displayed in Figure 5. In this data set, content about Black women was by far the most toxic on average. This mean toxicity in the Black women toxic data set was 0.73, which is 15% higher than for white women and 17% higher than for women in general. Furthermore, we found almost 40% more messages (28,728 in total) in the ‘Black toxic’ category than for the category ‘white toxic’ (21,681 in total) overall. This reflects the way abuse against Black women is racialised.

40% more messages in the ‘Black toxic’ category than for the category ‘white toxic’.

This reflects the way abuse against Black women is racialised.
FIGURE 5. TOXICITY OF MESSAGES COLLECTED USING SLURS

WOMEN IN GENERAL

WHITE WOMEN

BLACK WOMEN

98,694 (29.6%)

7,290 (33.6%)

20,091 (69.9%)

39,605 (11.9%)

2,969 (13.7%)

1,293 (4.5%)

85,104 (25.5%)

5,266 (24.3%)

1,779 (6.2%)

52,800 (15.9%)

1,832 (8.4%)

3,940 (13.7%)

56,889 (17.1%)

4,324 (19.9%)

Highly toxic

>0.8

Least toxic

(Lightest shade)
Furthermore, the data collected using toxic slurs for Black women included by far the most content with the highest toxicity rating: **69.9% of the posts in the ‘Black toxic’ were in the highest category of toxicity, compared to 33.6% of the ‘white toxic’ data set.** We compared this data set to the baseline data of randomly collected posts from each platform. For the data set that refers to Black women with slurs, the average toxicity is **10 to 18 times higher than the average data used for a baseline comparison.** This vividly demonstrates the deeply harmful nature of online misogynoir, which has “life threatening repercussions for Black women and girls especially Black queer, and Black trans women and girls. One consequence is violent extremism. Black women are discredited, mocked, and disrespected on social media platforms.”

We also collected a second data set using neutral keywords such as ‘Black woman’, ‘white woman’, ‘Black girls’, ‘white girls’, ‘women’, ‘girls’ etc. This led to a surprising finding: as is shown in Figure 6, the toxicity for the ‘Black neutral’ category is lower than that of the ‘white neutral’ category and the neutral category for women in general. There are several reasons for this unexpected finding. First, we realised that much of the toxicity in the ‘white neutral’ category actually related to other races; for example, posts that criticised white women for being in mixed race relationships (e.g. ‘m*dshark’). A second important piece of context for understanding this finding is that there are many more commonly known and more specific racial and/or sexist slurs (e.g. ‘n*gress’ or ‘sh*boon’) for Black women than for white women. When Black women are referred to in a negative way, these slurs will usually be used (meaning these posts are part of the ‘Black toxic’ data set). When Black women are referred to with neutral terms, it will also often concern a message of neutral or positive nature. There are very few commonly known slurs for white women, other than the phrase ‘Karen’ – which is highly contested. Although ‘Karen’ is used as a derogatory term, it is used as a label for someone who weaponises their white womanhood to terrorise Black people, and therefore criticises a behaviour rather than an identity. Following a Black feminist analysis, it is crucial to consider power dynamics when evaluating hate speech. As a result, and also due to the high number of posts which referred to women actually named Karen, we decided to exclude the word ‘Karen’ from the initial data collection.

When people refer to white women in a negative way, they more often do so without specific slurs and more by using a neutral term for white women in combination with words that express a general feeling of hatred or displeasure (e.g. ‘degenerate white woman’ or ‘all white women are vile’). In our data set, in highly toxic messages (i.e. those with a toxicity score over 0.8), white women are more often referred to with neutral terms, while highly toxic messages about Black women most often use slurs specifically targeted at Black women.

Lastly, there is also a large online community supporting Black women and creating online safe places for Black women precisely because they endure online attacks. For example, many Black women have created pages or groups specifically for Black women, or for creating awareness around either the skills or the struggles of Black women as Moya Bailey describes in her book. This community consists largely, although not exclusively, of Black women. In these spaces, Black women discussed the dehumanising comments they encountered online. As a result, many Black women seek refuge in niche online communities. This indicates that groups who are more likely to be subjected to online abuse can feel compelled to surround themselves ‘digitally’ with the right people to help protect their wellbeing – as they may do offline.
Therefore, gathering data using neutral keywords to refer to Black women (e.g. ‘Black women’, ‘Black girls’), will also result in a lot of positive messages, which decreases the mean toxicity. This will be less so for white women and women in general. This is supported by running a search for posts which used positive words like ‘support’, ‘love’, ‘courage’, ‘respect’, and ‘admire’ (see Figure 6).

When we ran a simple topic model on these messages, we found that support for Black women was reflected in the topics (e.g. ‘Black maternal health’, ‘Black woman book club’ and ‘Black women entrepreneurs’). This shows evidence that Black women have been actively creating their own safe spaces to support Black women. While the effect is powerful enough to bring down the mean toxicity of the Black neutral data set far below the white neutral and the women neutral data sets, it does not alleviate the harm caused by online abuse. Mean toxicity is the most simple and clear statistic to use. However, it is important to remember that focusing on mean data can disguise a lot of important nuances and information.

Despite how toxic misogynoir is on platforms, we see the power and joy of Black women’s online communities and how social media is used to facilitate important conversations. This is why improvements to content moderation practices and tech accountability are needed in order to safeguard users using platforms, as intended – a topic we return to in our Calls to Action.
“I found a community online from a super early age. There weren’t many African people around while I was growing up. We were (according to my dad) the second Sudanese family in Brisbane. I grew up in a very homogenous, monocultural society. For me, going online was about discovering and gaining access to the world. ‘I learned so much. The conversations around race and social justice I had access to online were worlds away from what I had around me physically. The online space was always a place that felt like home in a way. I felt I belonged, that I had rights, and I found myself invested. The relationships I created online were exciting and were very real to me.”

-Yassmin Abdel-Magied
Returning to the ‘Black toxic’ data set, which contained the highest proportion of toxicity in our sample, we ran a topic modelling algorithm on posts in this category. Figure 7 shows the distribution of the data per topic in the ‘Black toxic’ data set, identified using topic modelling.

As we discussed in the introduction, racist tropes such as ‘the angry Black woman’ (‘Sapphire’), fetishisation (‘Jezebel’), and fatphobia (‘Mammy’) have underpinned the dehumanisation of Black women long before the Internet. A long tradition of Black feminist theorists have critiqued these ‘controlling images’ for the way they delimit Black women’s possibilities in the world, as well as the ways they inform racist policies. Our automated topic modelling identified several topics that resonate with these racist tropes (see Figure 7, which shows the frequency of these tropes in the ‘Black toxic’ data set):

**Angry “Sapphire”**

**Criminal**

**Ugly**

**Dumb “Mammy”**

**Monkey**

**Promiscuous “Jezebel”**

This section will include references to explicit sexist and racist terms.
This section goes further in depth on the ‘angry’, ‘ugly’ and ‘promiscuous’ tropes. We examine how the data we collected deploys these tropes, and explains why they are harmful.

1. The Saphhire or “angry black woman” trope

This was the most common trope referenced in the ‘Black toxic’ data set, with over three thousand instances. This misogynoir trope is hugely prevalent in our society, often due to how Black women have been consistently portrayed in popular culture such as film and art. This trope presents Black women as “rude, loud, malicious, stubborn, and overbearing”. It creates the extremely harmful and violent idea that Black women speaking up on matters that are important to them, no matter how mundane or how significant, are not an example of wanting to “improve things; rather, she criticises because she is unendingly bitter and wishes that unhappiness on others”. This often leads to tone policing of Black women’s voices, as well as dismissals and/or criticisms of their contributions to ‘public’ discussion, including on social media. This is even more prevalent when Black women speak out on misogynoir itself, which allows the abuse and harm caused by this form of hate to be further perpetuated as Black women speak up against it. The prevalence of this trope demonstrates the dehumanising nature of online abuse that erodes Black women’s right to freedom of speech and self expression.

2. Body-shaming tropes

Body-shaming posts denigrate the bodies of Black women and label them as inferior. The most common insults target Black women’s weight. Fatphobia is part of an oppressive tool to keep Black women away from their bodily autonomy. Derogatory comments are made about hair, make up, body hygiene, body shape and facial characteristics. Two core themes within body-shaming misogynoir are repeated through-out this clustering of data:
3. Masculinisation of Black women:

There is a long-standing tradition within misogynoir abuse of using masculinisation as a way to denigrate and dehumanise Black women, creating a division between ‘women’ and ‘Black women’ who are othered as not fitting into society’s definitions of womanhood. It links to body-shaming and racist perceptions of beauty as:

“the Black standard of beauty tends to be light-skinned, with loose curls and eurocentric features. A woman must have the body of a Black woman but the face of a European woman. Otherwise, her femininity and beauty are brought into question and scrutinised. Black women who do not adhere to this beauty standard are often masculinised and stripped of their femininity. Once a woman does not fit within the standard set to her, she is denied her femininity or respect.”

4. Racist comparisons of Black women to apes:

The origin of this form of misogynoir dates back to 18th century white European eugenics and supremacist ideas, sometimes referred to as ‘scientific racism’. It is based on the deeply ingrained ways white people can see themselves as superior and ‘the most evolved’ of the human race, leading them to abuse and denigrate Black people as the ‘least evolved’, and therefore closest to apes.

Both these themes, alongside other body-shaming content targeting Black women, contribute to the overall dehumanisation and increased risk of violence towards Black women online and offline.

5. Jezebel trope & fetishization

The ‘Jezebel’ trope and fetishization is a topic that was automatically identified by the AI tool by identifying clusters of messages that fetishise Black women. This topic is sexually explicit, derogatory and hypersexualises Black women in ways that are harmful and dangerous.
Fetishistic stereotypes are discussed, such as ‘big booty’ and ‘bbc’ which is a demeaning abbreviation for ‘big Black c**k’. Men indicate that they explicitly want to have sex with a Black woman, often because Black women are hypersexualised and therefore said to be more promiscuous. A lot of sexually explicit language in relation to Black women uses language linked to food such as ‘chocolate’ – further dehumanising Black women as they are described as something inanimate, to consume. Black women are also targeted within a discussion of heteronormative patriarchy about the ‘best quality’ women. This includes comparing women from different racial backgrounds and nationalities.

Generally, white women are stereotyped as the standard for purity and femininity; it is the white man’s burden to protect them from ‘beasts’ e.g. Black and brown men. Black women are often masculinised or depicted as angry or aggressive, however, the ‘Mammy’ trope is a slight departure from this norm. Within this, Black women are labelled ‘traditional’ and seen as partners who can provide docile support, linking back to misogynoir tropes of the ‘Mammy’ who is a reliable source of docile labour and support. This is discussed in opposition to white women who are labelled as ‘degenerate’ and seen to be ruined by western society due to ideas of feminism and empowerment. Women who are non-white and non-western are, in the context of the ‘Mammy’ trope, preferred as docile and feminine partners for men in line with misogynist ideas of women’s role in a relationship. This is shared alongside deeply racist abuse. There is a strong presence of fetishization of Asian women, particularly on 4chan, where they are labelled as the better alternative to western women in a continued context of racist and fetish content about Asian women.

We understand the harm created by this form of misogynoir as:

“Racial fetishization continues [a] pattern of cultural and racial essentialising in efforts to control black female bodies and sexuality. Racial fetishization is an extension and reproduction of white supremacist colonial racism. While it has manifested itself in the perceived glorification of black bodies, it is in practice incredibly problematic, since it prevents fostering a culture that permits black women to define their own sexual agency. Racial fetishism also contributes to [...rape culture, which tolerates and even validates sexual violence perpetrated upon black female bodies.”

69
Methods + Mastery

THE HYPERSEXUALISATION & FETISHISATION OF BLACK WOMEN FUELS ONLINE ABUSE.

Black women voiced how they were frequently objectified to feel like an “experience” that someone, usually a white man, can cross off a list of sexual fantasies. They also stressed how dehumanising it feels for them to face such “radically sexualised comments” including comparisons to food and other inanimate objects. Notably, Black women shared how they had encountered “racist tirades” after declining dates or sexual advances from men on dating apps, and sought advice from fellow victims online on how to best handle those kinds of interactions.

As part of this, Black women turned to one another to discuss tactics to protect themselves from these types of messages, comments and “microaggressions.”
Toxic tropes about Black women – whether they are ‘angry’ or submissive, ‘ugly’ or fetishised – contribute to the ongoing dehumanisation of Black women, not just in vitriolic speech online but also in racist policies and offline violence. These tropes have been identified and critiqued for decades, yet they are now proliferating in more extreme forms online. These toxic tropes are spread abundantly on both alternative platforms (and are fitted into their white supremacist/ extreme ideologies) and on mainstream platforms by users who are not necessarily connected to the alt-right or similar, more extreme groups. On mainstream media we mostly find stereotyping, body-shaming and fetishizing, while on alternative platforms, we find more white supremacist and antisemitic themes, which will be discussed further in the next section.

Posts related to misogynoir were often also linked to other forms of hate, such as homophobia, transphobia, and antisemitism. In other words, online communities and users who are dehumanising one group are likely to also be hateful towards other marginalised groups. These forms of hatred also reinforce harmful narratives of white supremacy, including conspiracy theories like ‘great replacement theory’. Hateful jargon and rhetoric of this kind is steadily trickling from alternative platforms to mainstream platforms in what can be described as a process of radicalisation.

HOMOPHOBIA AND TRANSPHOBIA

Clusters of messages that target Black women with homophobic and transphobic abuse included the fetishisation of Black LGBTQ+ women and gender minorities, as well as direct abuse and ridicule directed at them. This is expressed through explicit slurs, body-shaming and harmful stereotypes.

The data on transphobia stands out in this data. In particular, we see how transphobic slurs are used to shame and denigrate both trans and cis Black women. While Black trans women are demeaned as a caricature of ‘diversity’ and denied their basic humanity, cis Black women are targeted with
transphobic slurs as a form of masculinisation – linking to the body-shaming misogynoir described above.

On alternative platforms (Gab and 4chan), in particular, the data includes more discussion of the race of trans women, with conspiracy theories that there are more Black and racialised trans women than white trans women and suggestions that trans women’s whiteness means they should be ‘tolerated’ (in contrast to Black trans women), which related to the previously discussed masculinisation of Black women. Whereas on mainstream platforms, content includes more discussions of racism and allyship within the trans community, between Black trans people and white trans people.

A significant part of this data includes anger and fear directed at too much attention on diversity and inclusion efforts, particularly in the representation of Black LGBQT+ people. This is positioned as a threat to white society and much of this abuse is seen alongside content that mocks Black LGBQT+ women.

ANTISEMITISM AND RELIGIOUS HATE

This topic was automatically identified by the AI tool by identifying clusters of messages that target Black women with antisemitism, Islamophobia and other forms of religious hate.

Much of this hate is fuelled by conspiracy theories of an ‘evil Jewish elite’ pushing a ‘woke’ or ‘anti-white’ agenda and using white women and people of colour as tools for this cause. These included various posts related to the Covid-19 pandemic and the moon landing being fake. The antisemitism targeting both Black women and white women is disproportionately found on alternative platforms, although it does exist on mainstream platforms also. This resonates with existing research on how antisemitism and anti-Blackness overlap and reinforce each other, becoming increasingly conflated in a racist discourse from the mid-nineteenth century to the period of the Third Reich onwards.71

Black women are also targeted with Islamophobic abuse, particularly when related to wearing a hijab. Black Muslim women are described as ‘extraordinary evil’, demonstrating dangerous narratives that feed hate and violence against Black women.

Black Muslim women are described as ‘extraordinary evil’
WHITE SUPREMACY, RACE MIXING, AND GREAT REPLACEMENT THEORY

Topic modelling also identified a cluster of messages that target Black women with abuse related to white supremacist ‘replacement theory’ and related ideas of race mixing. Notions of ‘race-mixing’ were prevalent in (among others) Nazi ideology, which led to state policies of anti-miscegenation (prohibiting sexual relations or marriage between races) to protect racial ‘purity’.72

Online discussions in our data set focused on the ‘purity’ of the white race from a white supremacist perspective, meaning white people who are in relationships with Black women are described as ‘race traitors’. Harmful and toxic narratives of the dangers of ‘race mixing’ are spread, with specific reference to Black people being ‘inferior’ as per white supremacist notions of superiority to other races. This content links to notions of ‘anti-white’ politics and other topics that link to race (theory), race hatred and white supremacist ideology.

Harmful and toxic discussions of Black women focuses on their potential as romantic partners, relating back to white supremacist ideology. One narrative is that Black women are ‘unacceptable’ romantic partners due to race mixing, whereas another sees dating Black women and particularly having children with Black women, as a way of ‘whitening her bloodline’. This deeply dangerous notion is referred to as ‘bleaching’. Even abuse aimed at white women is often based on demeaning other races e.g. through racist vitriol against mixed race couples, in which white women are seen as ‘betraying the white race’. These posts were often linked to the ‘great replacement theory’ (see Key Terms).

The themes of this topic directly relate to the most severely dangerous content within the topic of violent and sexually violent threats explored in the section on tropes above. This link is made as the conspiracy theories and harmful tropes of Black men and women being more ‘criminal’ and/or prone to being violent, alongside fears of the great replacement73, justify the most violent threats and toxic abuse directed at Black women and men. The overall fear that ‘Black culture’ will take over ‘white culture’ is found in the topic related to severe misogynoir threats. This is related to broader fears that Black people will treat white people the way that white people treated Black and brown people in processes of colonisation, enslavement, and other forms of domination.

Many have highlighted the misogynist underpinnings of great replacement theory, which is heavily invested in controlling women’s bodies and reproductive rights.74 In this report, we also highlight how this conspiracy theory channels the same misogynoir tropes which have been used to justify threats of forced sterilisation or mandatory abortions against Black women.

FROM ALTERNATIVES TO THE MAINSTREAM

It is often assumed that this kind of extreme content is only shared on obscure, alternative platforms such as 4pleb (4chan) and Gab. While these alternative platforms are definitely still the most toxic, their content is increasingly finding its way to mainstream platforms.

In particular, hateful jargon from alternative platforms like ‘gorillion’ and ‘globohomo’ – and conspiracy theories, like the ‘great replacement’, are steadily trickling into the mainstream. Their existence on mainstream platforms, in turn, helps to normalise these ideas in everyday life.
For instance, the measuring of the use of specific, alternative platform jargon on mainstream platforms. Toxic content on alternative media platforms uses very specific code words, or ‘dog whistles’, and has a vocabulary of its own. We measured the use of this kind of jargon on Twitter. We collected tweets that contain keywords from a set that is specific to alternative platforms. Some of these words are specifically antisemitic, or white supremacist in nature. In addition, we also collected tweets containing words from a random set. Subsequently, we plotted the number of tweets using these words over a period of 6 years (between 2017-2023).

In Figure 8 below, the baseline (purple line) indicates the average number of tweets that use words from the random set over those 6 years. We see that this line remains fairly stable. The other lines show the number of tweets that use words typical of alternative platform jargon. For all of these words, we see a significant increase over these 6 years.

We compared the number of posts using specific keywords in the first half of 2017 to the second half of 2022. Again, we see a significantly greater increase in tweets using alternative platform jargon than in baseline tweets. Especially in 4chan specific slang (‘soyboy’, ‘globohomo’, ‘coomer’) we see a huge increase. But the use of antisemitic and white supremacist slang has also increased on Twitter (‘goyim’, ‘gorillion’, ‘great replacement’, ‘superior race’).

FIGURE 8: USE OF ALTERNATIVE PLATFORM JARGON ON TWITTER

* This set consisted of names of exotic birds and names of Dutch influential people (that have not seen a specific drop or raise in media attention). Keywords of which we expect no significant increase or decrease in tweets mentioning them.
At the baseline, we measure only a slight increase (14.19%). The graphs show that the use of alternative media jargon is also increasing on mainstream platforms such as Twitter. This includes antisemitic and white supremacist slang. This may indicate that the ideologies involved are also becoming more widespread on mainstream platforms and thus in our day-to-day society. Writing on this topic, Cindy Ma highlights how right-wing social media personalities bring pro-white identity politics into the mainstream while strategically evading accusations of racism.76 Existing cross-platform studies have shown that mainstream social media platforms – YouTube, Facebook, Twitter – have been some of the most effective at disseminating racist, Islamophobic, misogynistic, and transphobic ideas.77

Black women are too often sidelined in this issue and, as Alexandria Onuoha (2021) argues, it’s long past time for the field of far-right studies to make more space to discuss the role of misogynoir in digital media that influences violent extremism and the white consumption of misogynoir.78 For example, Arbeit et al provide a theoretical integration of research from various disciplines to create a framework for youth practitioners to counter fascist recruitment of white male youth.79

In a sense, the internet is revealing to us already existing forms of hatred. Digital misogynoir is an online manifestation of the racist and sexist system we all operate in. Furthermore, it is also part of a pattern of online radicalisation that is allowing those with extreme ideologies to share them broadly, connecting various forms of hate with each other in complex conspiracy theories. As Laura Bates, writer and founder of the Everyday Sexism Project, has explained, online spaces are the perfect home for all sorts of toxic and damaging behaviours and perspectives: “[the internet] provides ideal conditions for abuse that aren’t available in the physical world: […] an echo chamber where people with already problematic views can be radicalised very quickly and have their views confirmed and exacerbated by meeting a like-minded community in a way that wouldn’t have been possible at such scale before. There’s a deliberate radicalisation and grooming of boys – teenage white boys, in particular – into forms of white supremacy and male supremacy, which they might not otherwise have encountered. The impunity exists online and normalised, and therefore exacerbates, these kinds of behaviours.”80

By dehumanising the most marginalised and easily recognisable groups of people – such as Black women, queer people, and Jewish people – these forms of online toxicity spread a deeply dangerous and hateful vision of the world. Diverse forms of toxicity start to intertwine in one big cesspit of racism, sexism, antisemitism, and conspiracies, creating hateful rhetoric that is increasingly accessible in mainstream society. Furthermore, as Jesse Daniels writes, “white supremacy online sometimes leads to violence, harassment, intimidation, and racial terror, transcending the virtual world to damage real, live human beings.”81

Wow, that was pretty heavy! The previous pages included a lot of toxic and harmful content and subject matters. This is a reminder to breathe and recentre.
CALLS TO ACTION

To dismantle (digital) misogynoir, we need to take action both from the top down (i.e. institutional) and from the bottom up (i.e. grassroots).

Here are our calls to action for four groups of actors:

1. Tech companies
2. Governments and intergovernmental organisations (INGOs)
3. Research and civil society organisations
4. Online communities & digital citizens

The fact that everyone must play a part to keep our roads safe is now ingrained in our culture – whether that’s designating a sober driver ahead of a night out, or passing and complying with seatbelt safety legislation. The same attitudinal and behavioural shift needs to happen for online safety: we can all contribute to changing the current nature and reducing the scale and effect of online abuse. 

*We must collectively be committed to reducing and ending digital misogynoir.*

The work of ending digital misogynoir should not fall on those worst affected, i.e. Black women. Men and white people can and should be allies, although it is important to listen and understand what Black women have written on the subject of allyship, to avoid familiar pitfalls like ‘white saviorism’. We can all do better to detect and respond to misogynoir, whether as friends, family members, employers, within our communities, and in our jobs, schools and institutions.
CALL TO ACTION

TECH COMPANIES

Tech companies invented neither patriarchy, nor white supremacy, but they are responsible for the scale and spread of its online effects (and the harm it causes) at life-threatening speeds across their platforms. All tech companies must make a significant, meaningful commitment to diversity and inclusion – internally and externally.

Tech companies must include a clear definition of, and policies on, misogynoir, misogyny, racism and white supremacy used within their work on content moderation and online gender-based violence prevention. This will require tech companies to understand that misogynoir plays a key role in radicalisation into white supremacy, and that this is an urgent problem. As Cindy Ma argues, as long as the public, and tech platforms, hold outmoded ideas about what white supremacy looks like, new actors like the alt-right and alt-lite are unlikely to be held accountable for their dehumanising portrayals of minoritised groups such as Black women. Major platforms with resources to do so should include addressing radicalisation into misogyny and white supremacy into their task forces.

We are currently at a crossroads where tech companies face huge regulatory and legislative changes across different nations and jurisdictions at a rate never seen before, including new laws emerging in the European Union and the UK. While these laws differ in content, tech companies can choose to enhance their safety practices and make voluntary commitments, rather than waiting for implementation of specific legislative frameworks. With this in mind, we highlight some key areas below.

While Glitch has called for a mandatory ‘Code of Practice for Violence Against Women and Girls’ that takes a comprehensive, systematic approach to online gender-based violence through the Online Safety Bill, we call on platforms to voluntarily adopt a similar code of practice, where this is not a legal requirement.
Crucially, **platforms should carry out risk assessments** for how new and existing features can be negatively used against marginalised communities; these risk assessments should be made available for feedback and scrutiny from civil society groups and members of said groups.

Tech companies should also **release transparency reports with disaggregated data** available for independent research on content moderation: **including race, gender, and other identifying characteristics of abusive posts, flagged posts, and content takedown**. We are concerned not only about the lack of takedown when harmful content clearly violates platform’s own standards and the risks related to biased algorithms not being nuanced enough, but also by the way that content take down and moderation anecdotally appears to censor and suppress content in relation to women, members of the LGBTQ+ community and Black people.86 As Dr. Moya Bailey wrote,

> “more statistical data can become disaggregated to make those on the margins of the margins visible.”

Currently, some tech companies release transparency reports, but they do not collect or release disaggregated data.

We call on companies to announce a commitment to collecting disaggregated data in 2023, start collecting this data in 2024, and release disaggregated data by 2025 on these specific characteristics so we can have research and analysis on the entire content moderation process, from keywords that are deplatformed, AI tools used to takedown content, how much is human verified/takedown, as well as details on the appeals and wider enforcement tools. In particular, **we are interested to see the difference between AI captured content and what users have to self report**. In user reports, we want to know more about **percentages of victims vs bystanders reporting** (see further details in Appendix 1 on Transparency Reporting).
Tech companies need to stop clamping down on tools like Crowdtangle and APIs which allow researchers to collect social media data. Due to the nature of our methodology and the ongoing trend towards tech companies clamping down on APIs, the data we collected over the course of July 2022–January 2023 would not currently be available to us if we conducted the research a year later.

Tech companies should also invest in independent research on their products that are made publicly available, cooperating with anti-racist, LGBTQ+ and women’s advocacy organisations to fund research and projects to better understand the impact of misogynoir and other gendered and intersectional forms of online gender-based violence.

There have been clear and consistent calls for huge improvements to reporting mechanisms, with response times and easier approaches needed – for example, reporting multiple problematic posts at once is a frequently raised suggestion amongst those reporting online abuse, either as the target of the abuse or as an active bystander. Automated responses often do not go far enough.

We also need consistency in data set access across platforms. We need access to image and video content in order to analyse and moderate deepfake generated content. Lastly, all platforms should commit to a well resourced Trust and Safety Council that has a key priority pillar on gender-based violence that cuts across themes of age, verification, privacy as well as harmful content.

To improve content moderation, collaborate with experts who can critique and improve methods used for abuse detection from a Black feminist standpoint. AI is a tool among many others: there is a need for transparency & human input to make content moderation successful. Tech companies need to include expertise from a variety of lived experiences in AI annotation, implementation and interpretation. For instance, the abuse detection method we used initially failed to make the distinction between content directed at white women and content directed at light skinned Black women, meaning a significant amount of content related to toxicity on the theme of colourism was included in the toxicity rating of ‘white women’. This demonstrates a practical gap in the development of AI tools because of a misunderstanding or lack of awareness of misogynoir and the value of a Black feminist approach that can identify and contextualise colourism separately from white women’s experiences.

Content moderation should also not be limited merely to removing posts, but also to promoting communities and positive connections. In our own data set, approximately 10% of the ‘toxic’ data actually represents discussions about these harmful tropes, rather than abuse based on them. This distinction is crucial for prevention and content moderation, so that we are able to protect spaces where Black women and others can discuss the impact of misogynoir using the language included in these tropes, while being protected from abuse itself.
Twitter expanded its Hateful Conduct Policy in 2018 with the introduction of a Dehumanisation Policy. This new set of rules aimed to prohibit “language that treats others as less than human”, denying their human nature or their human qualities. It included an extensive list of the identifiable groups that could be victimised. However, this policy only protects individuals when they have been dehumanised based on “a membership in an identifiable group”. Glitch highlighted a limitation in the policy that an issue arises when a person belongs to two or more groups and their membership to these groups is not clear due to the nature of the dehumanising insult. For example, Glitch pointed out that Twitter was failing to pick up abusive tweets when a photo of a dark skinned Black woman was put next to a picture of an animal, highlighting the platform’s weakness in picking up anti-Blackness, especially with imagery.

Glitch’s subsequent role on the Trust and Safety Council – until the council was dissolved in December 2022 – meant that we experienced first hand how this policy was implemented.

Further policies focusing on the experiences of specific, marginalised groups have been developed at Twitter – notably, the 2018 policy that explicitly stated that it was a violation of Twitter’s rules to purposefully call a trans person by the wrong name (‘deadnaming’) or pronouns. However, in 2023, under new ownership, the company updated its content moderation guidance regarding hateful content, removing the policy that prohibited the targeted deadnaming and misgendering of trans people. Twitter has been accused of ‘quietly’ removing this policy, which sets a worrying precedent for the rolling back of progressive policies in the future, further highlighting the vital role of strong, informed journalism and civil society in this area.

CASE STUDY:
Glitch advocacy on hateful conduct policy at Twitter
(see further details in Appendix 1 on Transparency Reporting).
CALL TO ACTION

GOVERNMENT AND INGOs

1 Governments and intergovernmental organisations (INGOs) such as the EU, UN and OECD should implement a public health approach to online gender-based violence, as opposed to one based on criminalisation. A public health approach means looking at the symptoms and causes of online abuse/harms, and focusing on education and prevention rather than punishment (i.e. a carceral approach). A public health approach in tackling online harms would range from providing guidance to employers on how to keep their employees safe online, publishing national guidance on digital safety, and investing ambitiously in digital citizenship education. In contrast, current recommendations around public approaches to online harm and abuse emphasise a punitive approach. While punitive approaches may be needed in some cases, we also want to bring the question of transformative justice into the equation and see how a public health approach can also rely on robust education initiatives and community engagement that look at how to be a good digital citizen.

2 Therefore, governments should respond to the spread of extremism through improved curriculums at all levels of education because ignorance creates a vacuum that fascism so neatly fills. This should include, and not be limited to, providing high quality and up-to-date digital citizenship workshops that increase participants’ understanding of their online behaviour, the effects these behaviours have and understanding of bystander interventions to support the online ecosystem. Such interventions support young people to become responsible digital citizens with the requisite tools to navigate the online world and tackle the rise in online abuse.
Policy-makers should integrate both the concept of misogynoir and online abuse into broader work around gender-based violence as well as radicalisation. By investing in civil society and academic research, policy makers can develop a better understanding of the links between forms of hate such as misogyny, misogynoir, incel rhetoric, and white supremacy. Governments should also provide better education on digital citizenship, including education around misogyny and misogynoir. We must respond to the spread of extremism through improved curriculums at all levels and to the wider public no longer engaged in formal education.

One possible way forward would be to implement a *tech tax*, following a polluter pays principle, e.g. to efficiently and effectively combat digital misogynoir and online abuse, 10% of the revenue generated through the taxation of tech companies should be pledged to civil society organisations to help fund their vital work to end online abuse, such as training on digital citizenship and online safety. For example, in the UK, this should be enacted through the existing Digital Services Tax of 2% on tech giants like Facebook, Google and Twitter, or through Ofcom’s role as regulator under the Online Safety Act regime.

Lastly, governments and INGOs should hold tech companies accountable on misogynoir, particularly through legislation on content moderation and transparency reports. Governments should be ensuring that any response to gender-based violence online also includes an intersectional analysis and addresses the particular risks faced by women and girls from marginalised and racialised communities. This would include passing new regulation and legislation that take a gendered and intersectional approach to online safety. This would ensure strong implementation of new laws in this area, such as the Online Safety Act in the UK, which should be strengthened with an updated gendered and intersectional approach to the subsequent regulatory regime. They should also include smaller platforms in online safety legislation, with a particular focus on platforms which act as a conduit to white supremacy.

Governments should invest in programmes that focus on the recruitment and retention of women in science, technology, engineering and maths, particularly Black and minoritised women, to address the gender and diversity imbalance within the UK tech sector. Lastly, technology and algorithmic bias should be included in human rights frameworks. For example, Glitch has joined the Article X Coalition, a group of digital organisations from across the globe advocating for inclusion of an Article on freedom from Algorithmic Bias in the forthcoming UN Declaration on the Rights of People of African Descent, as a founding member.
RESEARCH & CIVIL SOCIETY

1

Researchers across the tech industry, academia and civil society should work together to adopt an intersectional approach across sectors: abuse is not ‘just online’, ‘just misogyny’ or ‘just racism’. Everyone in the digital rights space needs to “be talking precisely, and helping to equip all women in politics, campaigning, journalism, blogging, sports and entertainment – not just white women. In particular, we call on those in the research space to fund Black women to do research degrees and to fund those already in research. We need to look through an inclusive lens when we’re tackling online abuse, so people aren’t being left behind.”91

Digital rights work needs Black feminist analysis, and work on gender-based violence needs to include intersectional and digital analysis. Researchers and civil society actors should also (continue to) focus on marginalised groups which are often lost/excluded from universalising online safety work. Intersectionality should be standard practice in online safety research.

In addition to influencers, politicians and other hyper visible people in the public eye, we need to focus on everyday minoritised people who don’t have the social capital to challenge their abuse.92 This must include all Black women (cis and trans) as well as people across the gender spectrum. While we prioritise Black trans women, we understand that some Black nonbinary, agender and gender-variant folks, as well as Black trans men are marginalised in the digital space because they may be perceived within a gender binary and therefore subjected to misogynoiristic harm.

While this research focuses on five platforms, misogynoir is prevalent elsewhere across the online space and further research into misogynoir is needed to highlight this. In research conducted by Methods + Mastery on Glitch’s behalf, they found that people recounted their abusive experiences across other platforms including Reddit and TikTok, as well as
more niche platforms like Discord and even Roblox.

Similarly, future research should focus on languages other than English: English text-based content moderation is by far the most invested worldwide across various platforms, yet we still see huge gaps, so what are the implications for other languages?

Lastly, we call on funders in research & civil society to fund research for, by and about Black women: fund Black women to do research degrees and fund those already in research. This kind of research is still too often dismissed as ‘too niche’ – which is itself a form of institutionalised misogynoir, given the extreme levels of risk and abuse Black women face.

CALL TO ACTION

COMMUNITIES

1 At the grassroots level, individuals and communities should work to understand the harmful effects of misogynoir, as well as the racist and sexist tropes that underpin it.

2 We can also practise good digital citizenship by challenging misogynoir and supporting Black women online, as well as contributing to their joy, rather than harming them. We can curate our feeds to follow, listen to and amplify Black women across various intersecting identities, building community with different people, understanding their experiences and leading from a place of empathy and compassion.

3 Lastly, we can demand better from tech companies – sign up to Glitch’s newsletter to join our upcoming actions.
CONCLUSION

Without safety mechanisms for Black women using social media, there is no online safety for all. It is clear from the evidence that policy recommendations and safety-by-design mechanisms that do not counter the prevalence of anti-Blackness on the platform, in content moderation and innovation are not safe for anyone. This research also makes it clear that ignoring abuse, toxicity, and harm against Black women means current efforts to stop radicalisation and hate online are missing a key part of the extremist journey, and platforms’ roles in the collective safety of groups targeted by extremist hate. Aside from finally plugging the gap in protections for Black women in online spaces, this report further evidences the critical need for online safety efforts to see Black women’s wellbeing as fundamental to broader, collective change. Gender and race should not be separate conversations when talking about content moderation, safety, policies or regulation but interwoven in all conversations, including broader topics of privacy, ethics and tech for good.

Unlike the vast majority of existing research on online hate, our report provides both an intersectional analysis and a cross platform comparison, which allowed us to document how hate is moving from the fringe platforms into the mainstream. Our findings show that misogynoir is prevalent across all platforms, and that Black women do get a disproportionate amount of the most violent forms of abuse. We also documented how sexist and racist tropes about Black women, identified and critiqued by Black feminist theorists, persist in new forms of social media harassment. Lastly, we show how misogynoir connects to forms of hate besides racism and sexism, including antisemitism, homophobia, transphobia and white supremacist conspiracy theories like the great replacement theory. These have serious implications for our democracy, free speech and wider human rights – impacting marginalised groups’ abilities to express themselves online freely and without harassment.

Our report looked only at text-based abuse. Yet, we know that content moderation is even harder for video and images – not to mention newer emerging technology like large language models. Given the poor track record of monitoring (relatively simple) text-based content, we must ask what platforms are doing to mitigate the risk of misogynoir and other forms of hate before they push these products into the market?

Tech companies can do so much more to address hate on their platforms, but they have made a deliberate choice to put profits over people. All digital platforms have known they have a problem with online abuse, misogyny, racism, and radicalisation into conspiracy theories for years now, not least because Black women have been raising the alarm on it since the 1990s. It is a choice to delay, ignore, and deny the problem. It is a choice not to take more concerted action against digital misogynoir and other forms of online hate.

A rise in online abuse is stifling freedom of expression and joy. It is causing harm on a mass scale, including harm offline. Online abuse silences Black women through the controlling of images of Black women, and the violence and systemic abuse associated with the normalisation and far-reaching dehumanisation of Black women. This delay and denial contributes to stagnating our experience of joy online. What would it look like to have content moderation based on joy? What would it look like to promote content that inhabits and attracts joy, supporting community development? Drawing on Black feminist theory, we can reimagine online safety as joy and community: safety is, after all, a bare minimum. As Seyi Akiwowo writes, “We should be able to flourish online and experience freedom. Let’s move from being able merely to survive online, and escalate the conversations to be about Black women and all women, as well as other marginalised communities thriving online.”

"
REFERENCES

44. Mutale Nkonde, “Article X Coalition Report #1: Online Hate against Black Women: An Example of How Algorithmic Bias Impacts People of African Descent” (AI for the people, June 2, 2023).
47. Paul Röttger, "Improving the Evaluation and Effectiveness of Hate Speech Detection Models" (DPhil Thesis on Social Data Science at the Oxford Internet Institute, 2023).


71. Tudor Parfitt, Hybrid Hate: Jews, Blacks, and the Question of Race, Oxford University Press (Oxford, New York: Oxford University Press, 2021). See also bell hooks’ essay on the need for solidarity between


84. Seyi Akiwowo, How to Stay Safe Online: A Digital Self-Care Toolkit for Developing Resilience and Allyship (Penguin UK, 2022), 78.


90. Mutale Nkonde, ”Article X Coalition Report #1: Online Hate against Black Women: An Example of How Algorithmic Bias Impacts People of African Descent” (AI for the people, June 2, 2023).


92. https://github.com/HKUST-KnowComp/MLMA_hate_speech
94. https://www.clarin.si/repository/xmlui/handle/11356/1462
APPENDIX 1

NOTES ON TRANSPARENCY REPORTING

The following includes some notes based on Glitch’s feedback for transparency reporting at Twitter.

What work should we continue? Which aspects of the transparency report are most helpful:
- Country comparisons
- Characteristic of bad faith actors and being able to filter/mute these out while Twitter takes actions
- Granularity around policy violations

What can we do better? Tell us how Twitter can improve its transparency reporting -- e.g. breakdown of data, visual display, descriptions, including additional metrics.
- Granularity of the type of abuse
- Granularity of the type of hateful conduct
- Dehumanisation policy extension to race
- Being able to copy and paste from webpage
- Granularity of the types of account taken down
- Granularity of how the accounts were taken down i.e. algorithmic detections, 3rd party reporting, personal reporting- would be good to show how effective bystander reporting is
- How long it takes accounts to be taken down
- Time period of when those tweets/accounts made a violation- was this a backlog of tweets accounts or incidences from 2019
- Gender-based violence online- more information about that
- Additional metrics: how many people are using block, mute and filter functions
- Additional metrics: misreporting- would be good to know if people understand how to report
Online toxicity encompasses rude, aggressive, demeaning and degrading attitudes and behavior on online platforms (in this report, social media platforms). This page shows posts in our datasets alongside their toxicity ratings.
The American Genocide Machine
The Murderer of the White Race

Write this in the stars and chisel it in stone:
“America’s police and military powers, along with her media, industry, religion and finance are the agents used to mix, overrun and exterminate the White race.”

1) Only 8% of the earth’s population is White, and only 2% is White female of child-bearing age or younger.

2) The envitable result of integration for the White race is genocide, as has been proven throughout history. Forced or encouraged racial integration is deliberate, malicious genocide!

The movies worship criminal n****s and meth cooks (you know what the fuck I am talking about). They encourage promiscuity and actively mock chastity, loyalty, and practically all morals. Even the “conservative” based and replied American actively enjoy Jewish filth like “Always Sunny” and “Sopranos” (hahahahha Jamal Ginsberg!!!!) or what ever have you boomer sitcoms. They love watching white men, father be emasculated while their intelligent but loving wife fix everything.

Their games consist of mindless violence, sex, drugs, the usual stuff they enjoy in real life. They love to pay $60 (with DLC, of course!) to play as mulatto n****s and women killing the evil bad white guy. Recently they also hate masculinity and view it as toxic, non-integral part of civilization......
Prior to this research, Textgain tested the accuracy of this algorithm by running the toxicity classification tool on several existing datasets containing toxic data. The toxicity rating system performs with an above 90% accuracy rate for Twitter\textsuperscript{107} and alternative platforms\textsuperscript{108}, but slightly lower for Facebook.\textsuperscript{109} This is mainly because the system seems to be biased towards longer texts and judges them to be more toxic than they sometimes are (due to the accumulation of toxic word use).

To measure accuracy in the toxicity detection algorithm on our dataset, two human coders independently reviewed a subset of 500 random posts from our data set, coding each post as “toxic” or “not toxic”. We then compared results between the two coders, finding an inter-coder reliability rate of 83.8%. This means both coders had the same code for 419 out of 500 posts. After discussion between the two coders, we came to a consensus on combined codes for the data set. We then compared these combined human-allocated codes with the toxicity scores allocated by the algorithm, using a score of .333 as a cutoff for toxicity (i.e. posts with a score above .333 were understood as “toxic”, and posts below .333 were understood as “not toxic”). The rate of agreement between combined human coders and the algorithm was 83.6%, which means they had the same code for 418 out of 500 posts. It also points to the subjective nature of determining abuse: there is no fully objective measure of toxicity, and there will always be a degree of disagreement among both people and algorithms as to how toxic certain messages are. Given this subjective element, the algorithm determined toxicity with a very similar rate of accuracy to human coders.

The table below shows a “confusion matrix” which gives more detailed information about agreement between the human coders and the toxicity detection algorithm. The left-hand side represents the human coders, while the top represents the algorithm. This shows that in the sample of 500, there were 281 “true positives” (posts which were coded as toxic by both human and algorithm) and 137 “true negatives” (posts which were coded as not toxic by both human and algorithm). There were also 34 “false positives” (posts which human testers identified as not toxic but the algorithm identified as negative) and 48 “false negatives” (posts which human testers identified as toxic but the algorithm identified as not negative). This means that where the algorithm made mistakes (compared to human coders), it was more likely to underestimate the number of toxic posts than to overestimate them. Therefore, the statistics in the report should be understood as likely underestimating the levels of toxicity in our sample.
### TABLE 1. Confusion matrix

<table>
<thead>
<tr>
<th></th>
<th>YES</th>
<th>NO</th>
</tr>
</thead>
<tbody>
<tr>
<td>YES</td>
<td>281</td>
<td>48</td>
</tr>
<tr>
<td>NO</td>
<td>34</td>
<td>137</td>
</tr>
</tbody>
</table>
If this research has moved you to action, taught you something new, or provided an evidence base for your work, please consider donating to support Glitch’s work and the implementation of the report recommendations.

DONATE

www.glitchcharity.co.uk