Prepared for **UK Parliament Public Bill Committee, Online Safety Bill**
Submitted by **Imran Ahmed, CEO, Center for Countering Digital Hate**

Introduction

1. Thank you for the opportunity to provide you with evidence on the Online Safety Bill. Through our work at the Center for Countering Digital Hate, we have developed a deep understanding of the online harm landscape - CCDH have advised UK, US and other governments on disinformation, on violent extremism and on how conspiracy theories can asymmetrically overwhelm fact-checking countermeasures and cause considerable real-world harm. Our research and advocacy work, showing repeated failures by social media companies and demonstrating the outcomes of their algorithms on our information ecosystem - systematically biasing it towards hate and misinformation - has evidenced the need for legislation that changes the fundamental business models and therefore behaviour of the platforms who profit from the spread of misinformation, disinformation, conspiracy theories and online hate by bad actors and by their own systems.

2. This is a complex but essential area to regulate, which we have been advocating for since our organisation was first formed in 2016. CCDH was set up to disrupt the work of malignant actors online and to accelerate the moment when Big Tech has to take responsibility for the fact that at the same time it is a source of connection for many of us, it has powered modern extremism, science-denial, hatred, bullying, terrorism and child sexual exploitation. In fact, as evidence has shown over time, it has been designed in a way that companies profit from the promotion of this problematic harmful content and allow bad actors to thrive.

3. CCDH has studied and actively disrupted the way anti-vaccine extremists, hate actors, climate change deniers, and misogynists weaponise platforms to spread lies and attack marginalised groups. What has remained consistent, across all these types of hate, is that platforms fail to act.

4. The failure of social media companies to act on known racist content connected with terrorism, misogyny and online hate is a violation of their own terms and conditions, the pledges made to an international community when the cameras were rolling, and the inherent dignity that the victims of tragedies like Buffalo were entitled to - the right to live safely in their communities and to be safe from extremist, racist terrorism.

5. The recent shooting Buffalo in the United States has been a stark reminder of just how important it is to hold those who facilitate the spread of conspiracies and extremism to account:
   - First, the bad actors who seek to dehumanise, polarise, and spread disinformation by weaponizing online spaces; and
   - Secondly, the bad platforms that reap record profits as they systematically fail to act on harmful content.

6. Misinformation, identity-based hate, and malignant use of digital spaces are not unique to any one country. The future of online harm and misinformation is a global story, which will require global collaboration.

7. We commend the United Kingdom and Her Majesty's Government for being ahead of the pack on tech reform – the historic Online Safety Bill ("**the Bill**") joins the European Union's landmark piece of legislation, the Digital Services Act, as a bold attempt to fundamentally change the rules of the game for these companies and incentivise a duty of care to protect users. Collectively, we can lift standards and safety online.

8. Our experience as an organisation suggests that three things are missing from existing powers globally:
   a. The power to compel transparency around algorithms (which select which content is amplified and which is not); enforcement of community standards (which rules are applied and how and when); and economics (where, when, by whom, and using which data, advertising, which makes up the bulk of revenues for social media platforms, is placed)
   b. The power to hold accountable social media platforms at an individual, community and national level for the impact of content they monetise
   c. The power to hold accountable social media executives for their conduct as administrators of platforms that hold enormous power over discourse not just in terms of content moderation, but also: the amplification of content, institutional and user experience design of the systems through which discourse occurs, and equity in user experience for marginalised communities.

9. We have analysed the provisions of the Bill using our new "STAR" framework, which we developed for analysing legislative efforts globally. The framework includes:
   - **S**afety by design
   - **T**ransparency of algorithms, rules enforcement and economics
   - **A**ccountability to democratic bodies
   - **R**esponsibility - of platforms and their senior executives.

10. All of these components need to be present to create really effective legislation and the change that we need.

About the Center for Countering Digital Hate

11. The Center for Countering Digital Hate (CCDH) is a not-for-profit NGO that seeks to disrupt the architecture of online hate and misinformation.

12. CCDH has been at the forefront of unmasking how online platforms and search engines drive radicalisation, online harm and misinformation. The Center's work combines both analysis and active disruption of these networks. We champion levers for change to increase the economic, political, and social costs of all parts of the infrastructure - the actors, systems, and culture - that support and profit from hate and misinformation (for example, climate change denial, sexual and reproductive health, anti-vaxx, antisemitism, and identity-based hate). We have included a summary of recent research in **Appendix A**.

13. CCDH is independent, is not affiliated to any political party and does not receive money from technology companies. We believe it is impossible to serve honestly and without fear as an industry watchdog against harms an industry produces if they also pay our salaries. We have offices in London and Washington D.C., and connections globally.

Key Issues

CONTEXT + RISKS

14. We know that Big Tech resists this regulation. For a large part of the past three decades, they have claimed that they are “neutral” in terms of harm. Subsequently, as evidence mounts from organisations like CCDH and whistleblowers like Frances Haugen, Big Tech have been compelled to acknowledge that there is a problem with their products and services, which is causing harm. They have commissioned internal research that proves the same. That the commercial decisions that they have made in the design of their products and lack of enforcement of their policies has amplified this harm. They advised Committee members in their oral evidence on the Bill that they have “fixed all their problems”. This could not be further from the truth. The fact is they have adopted Big Tobacco’s playbook: deny culpability, deflect blame and delay change.

15. And globally, Big Tech has been fighting reform each step of the way with their extensive lobbying spend and tactics.[1] This cynical lobbying approach was summed up rather acutely by Alex Powers, Director of Policy and Public Affairs at BT. He said:

> "I think we could listen to everyone's contributions and play a game of corporate lobbying bingo: 'unintended consequences', 'inadequate consultation', 'inflexibility', and so on. The reality is we've spent five years going through this process. Broadly speaking, this is totally the right approach. This isn't an attempt to regulate every single bit of content on the internet. It's about the conduct of the largest companies that affect everyone's lives from day to day."[2]

TRANSPARENCY + ACCOUNTABILITY

16. There are important principles stated in the Bill about, for example, privacy, freedom of expression, matters of democratic importance, and journalistic freedom. The Center supports these principles as fundamental democratic freedoms. We have also seen how these principles have atomistically been used to justify inaction (and underinvestment) by Big Tech who take, for example, a narrow approach to <u>whose</u> freedom of expression is being protected or prioritised and ignore the chilling impact that hate speech, misogyny and racism has on those who are subject to it without recourse. Instead of a platform for speech, it becomes a platform for bullying and extremism, and a misinformation highway network.

17. The Bill enables the regulator (in partnership with independent civil society, international regulators, and skilled persons under clause 88) to scrutinise the approach to these issues taken by Big Tech, particularly as it translates to deciding on, and implementing safety measures and policies.

18. In making this assessment above, it is worth emphasising that these companies are not, despite their vital role in public discourse, in the business of "free speech". They are motivated by money and make that money by:
    a. Selling users' personal data, and insights derived from that data;
    b. Selling advertising space on content users produce which they purport to hold to stated standards;
    c. Providing infrastructure - web hosting, monetisation, and customer relationship management - to other organisations seeking to access digital audiences.

[1] https://www.washingtonpost.com/technology/2022/05/17/american-edge-facebook-regulation/
[2] https://diginomica.com/online-safety-bill-google-facebook-tiktok-lawyers-respond

19. There needs to be clear legal requirements and effective enforcement and accountability to change corporate behaviour and embed safety by design - particularly when they have been allowed to flourish unchecked for so long.

20. We support greater clarity in the Bill to enable independent organisations like CCDH to be able to access data to analyse how well the platforms and search engines are meeting their obligations and to identify existing and emerging harms. OfCom will not have this capacity inhouse and these relationships with civil society independent from Big Tech are critical to the full ambition of this legislation being realised.

21. We were pleased to see that the Government has adopted our earlier recommendation enabling Ofcom to perform independent audits of platform safety - as per new Schedule 11, and that presumably this can be used in combination with clause 88 on expert reports so that Ofcom can rely on external capability and skills. Ofcom needs the ability to be able to act in these situations, though, as it relates to lawful content and reinstate the ability to mandate "proactive" tools for content moderation are used in relation to legal content, which was in the draft bill. In addition, there needs to be the ability for Ofcom to examine the platforms' systems and response to abuse in direct messages.

22. Clause 137 of the Bill sets out provisions relating to Ofcom's reports, including certain areas for exemption based on seriously or prejudicially affecting the interests of that body. Further thought needs to be given to whether to include an additional criteria of trumping this exception where there is an overriding public interest in favour of disclosure. There may be legitimate problems that are left unaddressed and there is a public health or other emergency that means public disclosure is beneficial. One example of this may be where a bad actor spreading COVID-19 disinformation is removed from the platform and loses advertising revenue and money as a result.

JOURNALISTIC CONTENT EXEMPTION

23. Like the exemption for content of democratic importance, the principle behind the protections for news publishers in the bill are absolutely correct.

24. We already have a press regulation regime in the UK. It does not make sense for publishers to be regulated twice - first by a press regulator or by Ofcom's broadcasting regulations, and again by the measures in the Online Safety Bill.

25. We should also remember that it is ultimately platforms that will be responsible for implementing many of the measures in the Bill, and we would not want a situation where platforms are unduly interfering with the operations of news publishers, especially when this is an industry they have already caused a lot of disruption to.

26. The problem here is that the Bill gives such a loose definition of "news publisher" that these protections could potentially apply to any website that meets conditions, like posting a complaints procedure and a set of standards.

27. Indeed, we have seen in our work that many websites spreading vicious racist hatred and dangerous misinformation already list supposed complaints procedures and standards on their websites simply to fool readers into thinking they are more legitimate publications.

28. There is no comparison between *The Times* and a blog that shares race hatred alongside Covid misinformation. The bill must be amended to protect legitimate news publishers while allowing platforms to act on content posted by sites whose entire business model is built on promoting racism or misinformation.

29. Like the democratic importance issue, we recommend that the Committee take a closer look at this clause to ensure that the regulator has the power to act in respect of small hate / extremist / disinformation sites who may purport to be subject to accreditation and standards but are outside the scope of comparable British regulation for journalists.

30. We recognise that there is a careful balance to strike here between supporting new forms of media and citizen journalism and genuine grassroots local journalism, and those sites that are designed and / or funded to spread harmful content. We recommend working with independent civil society and the journalism profession in these assessments. Bad actors should not be able to pass the test of "journalistic content". At the moment this clause is vulnerable to being tested by e.g. people who wish to spread misinformation for monetary gain. One option, in the first instance, may be

for companies to provide transparency reports to Ofcom about who was provided with legal protection under this exemption so that there is a mechanism for public scrutiny, civil society advocacy and some reputational moral hazard for social media companies to discourage a damagingly over-broad interpretation..

STATUTORY PRINCIPLES

31. We understand that there is some nervousness about what the detailed requirements will be in secondary legislation and the Codes. Conversely, in this fast moving environment, there also needs to be flexibility to ensure that the regulator is able to react quickly to emerging problematic types of content and steps that can be taken under a Code. There are a series of consultation requirements and Parliamentary scrutiny steps that need to be met in respect of these documents. It may be useful to amend the Bill to add in statutory principles, to guide how duties and decisions are made under the new legislation and to help future-proof the legal framework as technology develops.

MISINFORMATION + DISINFORMATION

32. Our understanding is that mis/disinformation could be brought into scope by being listed under **"priority content that is harmful to adults"**, as set out by the Secretary of State in secondary legislation, and that in some cases it may also fall into the unlawful category. In CCDH's evidence to the Joint Committee on the draft Bill, we called for misinformation to be put on the face of the Bill, and recommended that it also form part of a positive duty for platforms to "mitigate and manage" as a category of content that is harmful to adults. We believe that mis- and disinformation must be addressed, especially the type of disinformation repeatedly highlighted in our research which poses a direct threat to public safety and democracy.

33. We note that this concern was shared by the Joint Committee, who specifically referred to the need for electoral and misinformation / disinformation to be included as priority harms on the face of the Bill. The last two years have seen the full impact of COVID and electoral misinformation, which in a very real way has contributed to deaths from COVID-19 and extremist events, such as the unfounded Stop the Steal campaign escalating into an attack on the Capitol on January 6, 2021. Election and public health misinformation/disinformation are known harms. It is unconscionable that bad platforms and bad actors have benefitted commercially from the spread of this content, and it needs to be explicitly recognised as a harm to adults and incorporated into the framework of the Bill.

34. Further thought should also be given to the role and functions of the Advisory Body on Disinformation and Misinformation in clause 130 of the Bill, and how the recommendations from that Advisory Body fit with the duties on platforms, the codes of practice and powers of the Secretary of State (such as

setting principles) and other advice and reports that are received by the Secretary of State, Ofcom and Parliament.

SMALL SITES + BAD ACTORS

35. The general focus of the Bill is on the large platforms and search engines. Through our work, we have also identified a number of issues with small websites and platforms, and individual super-spreaders of online hate, misinformation and extremism. We have some specific recommendations related to these areas, which include addressing the economic drivers and reach of these bad actors and bad platforms.

36. Smaller sites are exempted from Category 1 duties by virtue of their size alone – there are no considerations for the degree of harm they cause. We recommended that they be included in the scope of the review of the regulatory regime for two years after commencement and that further thought should be given to enabling Ofcom to use their powers to disrupt their activities - e.g. business disruption and takedown notices.

37. Alongside this, the safety duties protecting adults are undermined by the ability for providers to include in their terms and conditions that their treatment of priority content that is harmful to adults is to be treated by "recommending or promoting the content" (cl13(4)(d)).

38. To provide a small insight into the problems from these smaller sites:
    - **Incels:** CCDH discovered that a tiny number of small forums provide a platform for radicalisation of young men in the UK. In 2021, just three sites combined received hundreds of thousands of views by UK users each month. This has led to offline violence.
    - **MP abuse:** CCDH uncovered that Telegram played host to closed groups in which extreme threats of violence were made against named UK MPs. Not only is there virtually no oversight or content moderation taking place on Telegram, it takes place behind closed doors.
    - **Anti-vaxxers:** Under the threat of being de-platformed by Big Tech platforms, anti-vaxxers "lifeboat" followers from the major platforms into smaller ones, where they can spread more extreme forms of misinformation and conspiracy theories.

39. It is worth noting that these smaller sites often generate traffic from content shared on the larger sites - and build revenue that way, so there is a connection between the reforms to the main platforms and the smaller sites. Online trolling remains an issue. For example, through our research in 2021, we found that anti-vaxxers who spread misinformation about vaccines and Covid were doxxing those dedicated to spreading factual information about

vaccines, and directing their hundreds of thousands of followers to troll and abuse them.

40. Ofcom will need to have the ability to act quickly when problems arise - for example, a website dedicated to spreading electoral misinformation may be established for a set period of time calculated to cause the most harm, and has the potential to wreck a lot of damage to the democratic process. Similarly, the ability for platforms to enforce policies against hate and misinformation bad actors, and specifically to deal with trolling behaviour, will be important.

41. In addition, we know that there are whole sites whose purpose is to propagate online hate and misinformation, many of which receive substantial profits from advertising revenues, from companies and organisations who have no idea that they are funding these groups. This issue and a solution for addressing it are explained in more detail in **Appendix B** to this paper.

OFCOM SCOPE + POWERS: ONLINE ADVERTISING

42. The Bill was amended from the draft version to include fraudulent advertisements in scope, which extends to duties on Category 1 and Category 2a providers. Ofcom needs to have a role in regulating all forms of online advertising. In general, with the introduction of the new regulatory regime in the Bill, Ofcom will develop a unique perspective on the dynamics and drivers of online harm from a regulator's perspective. The current self-regulatory model for general advertising is not appropriate for regulating this mode of harm - which is recognised in other clauses of the Bill, such as clause 77(5). The risk with not including paid advertisements in scope is that a person may easily turn their content into a paid advertisement to escape the regulation. Providers' systems should not discriminate between paid and unpaid content when they are looking at harm, particularly harm directed at children. We understand that there is a separate DCMS review underway on regulating advertisements more generally but think that online advertisements should be progressed as part of the online safety regime. We support 5 Rights' recommendations about

REVIEW OF THE LEGISLATIVE FRAMEWORK

43. Our focus is to ensure that this legislation is strengthened so that any attempts to thwart the process and public safety safeguards introduced by the Bill are minimised. There will be a constant and ongoing need to ensure that the legislative framework remains fit for purpose and responsive to preventing and responding to online harms. We have specific recommendations about extending clause 149 of the Bill, which relates to the

statutory review after two years of operation. Given the rate of change in this sector, further consideration needs to be given to the frequency of these reviews to ensure that the legislation meets its purpose on an ongoing basis. There is an opportunity to try / adapt new intervention tools and to adopt a continuous learning mindset for this legislative scheme, to learn from what has worked well. Having statutory principles for how decisions are made under the Act, connected to the statutory purpose, will also help to ensure longevity of the legal framework over time.

44. We recommend a comprehensive, independent review. This would mean extending the specific matters currently covered in the statutory review in clause 149 of the Bill to include, for example:
    - Scoping of emerging harms and convergence of harmful content;
    - Whether and what additional regulation and interventions are appropriate for problematic small websites and search engines; and
    - The impact of any new evidence, global standards, overseas law reform or commercial changes.

FUNDING OFCOM

45. While we understand that this is largely designed to be a cost recovery regime with fees imposed on relevant services, further thought should be given to whether the baseline funding for Ofcom is sufficient for establishing its new functions before full implementation, i.e. the Committee may like to clarify whether any fees recovered are able to be applied for these preliminary costs before the Secretary of State has made the principles and Ofcom has calculated the relevant amounts in the fees statements.

46. It is right that the Select Committee seeks further refinements and improvements as the Bill works its way through the Parliamentary process. We also recognise that with this Bill, we are breaking new ground, and that it is important to ensure that the resource and capability of Ofcom is going to be directed at the most egregious content and the platforms that have the most widespread impact on people in the UK - at least in the first instance. We do not want to see the regulator set up to fail - it is too important.

CYBERFLASHING OFFENCE

47. We support the new cyber flashing offence that has been introduced into the Bill. As an organisation, we have extensively researched online misogyny and abuse - including our most recent report Hidden Hate which analysed the Instagram DMs of five high profile women. We found Instagram failed to act in 90% of cases of serious abuse - including death threats and sexual abuse -

even when the content is reported using Instagram's own complaints mechanisms.

ADULT RISK ASSESSMENTS: MITIGATING FACTORS

48. Clause 12(5): Adults' risk assessments outlines a list of matters that need to be taken into account when making an adults' risk assessment. It explicitly refers to:

- "The harm that might be suffered by adults". This provision should be extended to include consideration of the offline actions that the person may be inspired to take which harm others - i.e. the broader social, criminal and environmental harms.

- Media literacy is accepted as a mitigating factor in clause 12(5)(h) but no specification of this being "independent" media literacy programmes. The integrity of this provision would be enhanced if it did not extend to inhouse programmes. Public safety will be enhanced with unbiased information, and running an inhouse programme is not sufficient as a factor to mitigate the identified risks from Big Tech's products and services.

Response to Questions from the Select Committee

In your view, what needs to be added to or taken away from the bill to help it achieve its stated aim of making the UK the safest place in the world to be online?

49. Refer to issues outlined above and responses below.

What do you think of the decision to remove misinformation and disinformation from the scope of the bill?

50. See response in the issues outlined above. We recommend that it be added back explicitly into the Bill and that the Advisory Body be integrated with other key parts of the Act.

The Bill contains duties to protect content of democratic importance. What is your view of these measures and their likely effectiveness?

51. The principle behind this duty is absolutely correct - platforms should absolutely consider the democratic importance of content when making moderation decisions.

52. But we know from our work that misinformation and disinformation on social media poses a real threat to elections and democracies around the world. We are an international organisation, and we have studied the real harms caused by online election disinformation in countries like the US. We saw websites like Gateway Pundit profit from Google Ads to the tune of over a million dollars while spreading election disinformation that has led to real world death threats sent to election officials, and contributed to the events of Jan 6th. We do not want this to happen in the UK.

53. The problem with the "democratic importance" duty is that it is framed negatively, about preventing platforms from removing content, rather than positively about addressing content that undermines elections. And this is concerning because it is the latter that has proven to be damaging in the real world.

54. Therefore, while it's right that platforms should consider the democratic importance of content when making moderation decisions, there should also be a positive duty on platforms to act on content that is designed and intended to undermine our democracy and our elections.

What about the Journalistic Exemption?

55. Like the exemption for content of democratic importance, the principle behind the protections for news publishers in the bill are absolutely correct.

56. We already have a press regulation regime in the UK. It does not make sense for publishers to be regulated twice - first by a press regulator or by Ofcom's broadcasting regulations, and again by the measures in the Online Safety Bill.

57. We should also remember that it is ultimately platforms that will be responsible for implementing many of the measures in the Bill, and we would not want a situation where platforms are unduly interfering with the operations of news publishers, especially when this is an industry they have already caused a lot of disruption to. But the problem here is that the bill gives such a loose definition of "news publisher" that these protections could potentially apply to any website that meets conditions like posting a complaints procedure and a set of standards.

58. Indeed, we have seen in our work that many websites spreading vicious racist hatred and dangerous misinformation already list supposed complaints procedures and standards on their websites simply to fool readers into thinking they are more legitimate publications.

59. There is no comparison between the Times of London and a blog that shares Tommy Robinson's racist screeds alongside Covid misinformation.. The bill must be amended to protect legitimate news publishers while allowing platforms to act on content posted by sites whose entire business model is built on promoting racism or misinformation.

60. Nobel Peace Prize winner and Editor of the publication *Rappler* made the following observation at our recent Global Summit to address Online Harms and Misinformation:

> "When it spreads, the things that would've been effective as an antidote in 2016 are no longer effective in 2022. The weaker we get, the more people believe we can't find the facts, can't find the information, then the harder it is to come back from the brink. **How do we get back to a place where checks and balances work? Our experts don't stand a chance. Journalists don't stand a chance. Especially when the incentive structure… goes to anger and hate for the widest distribution… but that doesn't create democracy. It creates outrage, which creates mob rule. Virtual, physical, it's the same."**

Does the Bill give Ofcom sufficient flexibility to regulate smaller, higher-risk platforms?

61. Smaller sites are exempted from Category 1 duties by virtue of their size alone – there are no considerations for the degree of harm they cause.

62. Just to give you some insight into the problems from these smaller sites:

- **Incels:** CCDH has discovered that a tiny number of small forums provide a platform for radicalisation of young men in the UK. In 2021, just three sites combined received hundreds of thousands of views by UK users each month. This has led to offline violence.
- **MP abuse:** we uncovered that Telegram played host to closed groups in which extreme threats of violence were made against named UK MPs. Not only is there virtually no oversight or content moderation taking place on Telegram, it takes place behind closed doors.
- **Anti-vaxxers:** Under the threat of being de-platformed by Big Tech platforms, anti-vaxxers "lifeboat" followers from the major platforms into smaller ones, where they can spread more extreme forms of misinformation and conspiracy theories.

63. It's worth noting that these smaller sites often generate traffic from content shared on the larger sites - and build revenue that way, so there is a connection between the reforms to the main platforms and the smaller sites.

How do you respond to those who say that the bill risks setting an unwitting precedent for non-democratic countries who would seek to restrict the freedom of expression of their citizens?

64. The bill appears to strike the right balance between legitimate freedom of speech and protection from harm, with important checks and balances on both sides through, for example, Parliamentary scrutiny, an appeals pathway for users who have content removed or are subject to adverse action by a company. The freedom of expression safeguards in the Bill include, for example:
    - A duty to have regard to the importance of protecting users' right to freedom of expression within the law.
    - A duty to have a complaints procedure which enables users to complain if platforms are not complying with this duty, if their content or their use of the site has been affected
    - A duty for platforms who take action against content that is legal but harmful to adults to clearly specify their policies in their terms of service and enforce them consistently
    - A super-complaint mechanism through which complaints can be made to the regulator if platforms are significantly adversely affecting the right to freedom of expression within the law of users of the services or members of the public

65. Human rights and natural justice have been baked in. Our bigger concern is that Big Tech will state that human rights and privacy considerations are a justification to continue a light touch approach to safety by design and content moderation. There is already a trend towards greater encryption. This would undermine the purpose of the Bill and effectiveness of the regulatory

regime. As noted above, it will be important for the independent regulator to utilise their oversight and audit powers effectively to ensure that Big Tech aren't treating this as a loophole from doing anything.

66. It's also worth noting that at the moment, freedom of expression is not being supported. Trolling and other forms of online abuse have a corrosive impact on the freedom of expression for those people who are subject to the abuse. This situation is aggravated further where platforms fail to act on abusive content - which we have found happens roughly at least 80% of the time across a number of types of abusive content (see our research on misogyny, anti-black racism, antisemitism and anti-Muslim hate). It becomes an unsafe place - both online and offline - triggering offline violence, as we've seen only too recently in Buffalo.

67. All of that said, there's an important diplomatic effort that the UK can make, to work with other countries to explain the rationale and mechanisms that have been introduced in the Online Safety Bill, and to work with other independent regulators to track progress.

Does the bill give due protections to user privacy?

68. Privacy interests are in a number of provisions in the Bill. As above, it will be important for the independent regulator to utilise their oversight and audit powers effectively to ensure that Big Tech aren't treating these clauses as a loophole from doing anything to support public safety or designing this into their products or services as a way of escaping regulation.

One of the key objectives of the legislation is to ensure a higher level of protection for children than for adults. In your view, does the bill achieve this aim?

69. The Bill rightly has a dedicated Part setting out greater protections for children.

70. CCDH has endorsed the Californian Age Appropriate Design Code Bill and we agree with 5 Rights analysis that the definition should be more flexible so that Ofcom can look at harms to children, wherever they occur. As the issues are global, it makes sense to find alignment where possible, if it lifts the protections for the public, particularly children.

71. In answering this, we also want to point to research that CCDH did recently on user safety in the Metaverse, including for kids. Our researchers found that users, including minors, are exposed to abusive behaviour every seven minutes. Such as:
    - Minors being exposed to graphic sexual content;

- Bullying, sexual harassment and abuse of other users, including minors;
- Minors being groomed to repeat racist slurs and extremist talking points; and
- Threats of violence and content mocking the 9/11 terror attacks.

72. We reported all of these incidents to Facebook using their web reporting tool when we found them. All of our reports about users who abused and harassed other users went unanswered. That is a 100% failure rate from Meta.

73. Our study revealed that Meta's Oculus platform essentially had no age verification controls. The only thing you needed to access harms hosted in the VRChat app was the possession of a Facebook account (which nominally has an age limit of 13 and above, but is often ignored).

74. In analysing the Bill, you would want to ensure that tech companies have actually implemented sufficient risk assessments and age verification tools, so that children cannot access harmful material.

75. There is nothing to stop a child picking up an Oculus headset and directly accessing dangerous spaces with virtually no age checks or anything impeding / slowing them down ("friction").

What should be added to or removed from the bill to improve how it protects children online?

76. Refer to the answer above and general recommendations for improving the bill listed early in this paper. The issue of age verification is an important one. This needs to be specifically addressed and referenced in the risk assessments. Currently age verification is only one example of age assurance cited in clause 11(14). If platforms like Meta are relying on their existing assurance mechanisms, like our experience with accessing Meta's Oculus platform - then all you need is a Facebook account (nominally an age of 13 years but frequently breached). These existing age assurance mechanisms are inadequate / non-existent. The Bill should be more specific about what is acceptable or not and age verification is the cleanest route for this. Additional approved age assurance mechanisms could be approved through secondary legislation or guidance issued by Ofcom. Any collection of data would need to be strictly held only for the purpose for which it was collected, e.g. not shared with third parties or for the purposes of advertising to children.

Should key, known risks of harm to children be set out on the face of the Bill? If so, which harms should be included?

77. Yes, in addition to the above, we would recommend that the following be added:
    - Pro-anorexia and pro-self harm content (which we have encountered on Instagram and TikTok where it is rife);
    - Sexual abuse and harassment, which we uncovered in the Metaverse (the VRChat app available in the Oculus platform);
    - Extremist ideologies (e.g. incel and other misogynistic ideologies). Typically, young men are drawn into these subcultures;
    - Racist and other forms of identity-based hate; and
    - Disinformation - which we know impacts everyone, including children.

Are there sufficient systems in place to promote the transparency, independence, and accountability of Ofcom?

78. Ofcom is subject to accountability mechanisms in other legislation, such as the Communications Act 2003, the Freedom of Information Act 2000 and the Public Bodies Act 2011.

79. We recommend strengthening the consultation requirements in the Bill so that there is a greater role for independent civil society in the process, rather than risking capture by Big Tech in the detailed design of this regulatory regime. For example, the requirement to consult on fees statements in clause 75(4) gives a broad discretion to OfCom to consult "such persons that they consider appropriate". This is intended to inform the decision as to whether "the fees required under section 71 are justifiable and proportionate having regard to the functions in respect of which they are imposed." This is obviously not a decision that should just be left to Big Tech - turkeys voting for Christmas - but the current drafting would not prohibit that situation, it would be up to Ofcom.

To what extent does the bill empower individual users to take responsibility for their own safety online?

80. The Bill is fundamentally right: to target systems on the major platforms and search engines. There are also a number of key provisions relating to complaints/report systems, transparency, and education.

81. For too long it has been accepted that Big Tech has no responsibility and this should be up to individual responsibility. But there is a very obvious power imbalance here - Big Tech have designed their platforms and search engines in a way that promotes highly engaging content for profit. Frequently this is harmful content that breaches their policies and the law. This content is being pushed at you, it is not a choice that you have opted into. And in the

majority of cases we have found that these same platforms fail to act when reports are made about inappropriate and unlawful content.

82. There are too few safety controls and there is a consistent failure to act when things go wrong. You're also not privy to what is happening online with the algorithm or other processes that have been adopted. That information is held by Big Tech. Nor are the resources in an individual's hands. Again, Big Tech is making money from your time and data. The moral obligation is on Big Tech. Now let's fix the law so it is too.

83. It should be safety first, not safety optional or safety missing. Safety starts with the companies designing their products and services so that they are safe, with accountability to an independent media regulator, and offences / economic consequences for non-compliance.

The bill mentions anonymity and pseudonymity only once. Should the bill take a clearer stance on the issue of online anonymity?

84. Anonymity is only part of the problem. In our research we have found that many people are perfectly happy to be abusive in their own names. For example, our research into racism directed at footballers indicated that a majority of abusive accounts are not anonymous.[3] In these cases, it was the failure of the platforms to react to the reports of abuse, and the systems that enabled strangers to send abuse, that was the real problem.

85. We note that there are people who are anonymous online because they fear reprisal from family or an oppressive regime, and are seeking to connect with like minded communities - such as LGBTQI+.

86. Platforms should be required to consider the harms arising from anonymity as part of their risk assessment, and if there are severe issues then be required to take measures to address it. Ofcom should have the flexibility to require / set standards, if there's a problem, instead of taking a blanket approach.

Should people be able to use the internet whilst remaining fully anonymous?

87. See response above. There are circumstances where people may have compelling reasons to be anonymous online, e.g. for members of persecuted groups who live under oppressive political regimes. Most people, though, are online because they want to engage with family and friends, not bots or inauthentic users.

[3] https://news.sky.com/story/euro-2020-why-is-it-so-difficult-to-track-down-racist-trolls-and-remove-hateful-messages-on-social-media-12358392

88. Using a safety by design approach, the Bill should have the setting of interacting with authentic users as the default option. People should have to consciously choose to opt into interacting with bots and inauthentic users. To avoid the risk of platforms using and abusing the collection of personal data[4] the authentication process needs to be done in a way that doesn't just increase the stock of information that platforms have on their users - understandable privacy concerns.

89. There may be technical issues with some technology, such as VR, where you cannot check who someone is with reference to their other online activity.

If users are required to disclose their identity, should it be publicly, or to the platform?

90. We don't think all users should be forced to disclose their identities. If they were, we don't think social media companies should hold that information.

I've quoted a lot of statistics that the Center for Countering Digital Hate have produced with regards to online abuse directed at individuals with particular characteristics. In the previous panel I mentioned that the vast majority of this is done via direct messaging, sometimes through an end to end encryption on platforms. What concerns do you have about this within the bill and do you think that the Bill adequately accounts for that type of abuse?

91. Abuse by direct message is an insidious and persistent form of abuse, where an abuser will seek to exploit the private nature of the communication without accountability or redress. As advised during our oral evidence to the Select Committee, the recent research that we did on online misogyny (*Hidden Hate*) found that Instagram was failing to act on 90% of misogynistic content sent via DM after it was reported to them using their own reporting mechanisms. We also found that abusive voice messages sent via DM were unable to be reported at all and that the "vanish DM" mode required a user to view abusive content in order to be able to report it using Instagram's systems.

92. We recommend that the Bill is strengthened so that there is a positive duty on the platform to ensure that they have a complaint pathway for all types of abusive content, which can be triggered without subjecting the person to further abuse. This is safety by design in practice.

Part of the issue that we're seeing is that regulated providers have to rely heavily on the use of AI to facilitate monitoring and to take down problematic

[4] https://www.independent.co.uk/tech/instagram-invasive-app-privacy-facebook-b1818453.html

content in order to comply with the bill. However, several stakeholders have said, the adequacy of the algorithmic moderation to recognize the nuances and subtleties required in order to actively and effectively take down this content. What more would you like to see in the bill to try and counteract that issue that's been arisen?

93. As mentioned in the oral testimony, the appropriate technical fix or mechanism will vary by platform and what is appropriate in any given circumstance is a matter for the provider to work out with OfCom, under the framework of the Bill. Given how unsafe these services and platforms are, they will need to invest time, money and people to fix their problems. Like other industries, the overriding obligation is for the providers to ensure that their services and products are safe - and that starts with safety by design. Safety by design is the first pillar in our STAR framework for analysing legislative efforts globally. Fundamentally this means designing systems so that they are safe at the front end rather than having to retrofit actions after the harm has occurred / is occurring. In this situation, there are known problems that will need to be addressed, but the framework and technology will continue to evolve with these principles in place.

94. As discussed earlier in this paper, the Bill has appeal pathways for people who believe that their content has been unfairly removed, and there are also specific duties in the Bill for providers to consider freedom of expression.

There are significant enforcement powers in the bill. But I just wondered whether our two witnesses here wanted to talk to whether they thought those enforcement powers were enough.

95. As advised during our oral evidence on the bill, having offences that target both the senior executives and the companies themselves is an important tool in the regulator's suite of options for ensuring compliance with the requirements - we recommend both of these aspects within our STAR legislative framework as core components of R - Responsibility. This is an effective model used, for example, within health and safety legislation to action internal accountability and culture change within a company.

96. There are a number of enforcement powers in the Bill that enable the regulator to inspect and audit providers in respect of their duties. As mentioned above, the looser obligations appear to be in respect of small sites, mis and dis-information, and lawful but harmful content. We recommend that these areas are addressed in the Bill.

I raised some of your stats with Meta (Facebook), when they were here, around reporting and the number of reports that are responded to, for example, and

they basically said, this is not true anymore. We're great, obviously, paraphrasing, could you please let us know if the reporting mechanism on major platforms, particularly Facebook is not completely fixed? Or if there are still lots and lots of issues with it?

97. This is the whole basis of our work. Every day our organisation is finding substantial evidence of problems with providers and their failure to act on online hate content, to address hate groups or superspreaders of hate and misinformation, to apply the safety mechanisms they state they have (such as failing to label Russian-state sponsored propaganda about the Ukraine War), to design new products with safety in mind (e.g. the Metaverse).

98. As noted above, there remain substantial problems with platform and search engines failure to act, sitting at approximately 80% across studies.

You raise the point there about the abuse that was directed at election officials in America and do you think it should almost be a standalone offence to send harmful or threatening communications to elected people? Through MPs or counsellors, or mayors or Police and Crime Commissioners, or possibly even election officials, so people who are involved in the democratic process at the risk that that abuse and threats could have on democracy?

99. One of the things that we found in the response to releasing our misogyny report is just how widespread these issues are experienced by women in different professions. The women in the study were women with large Instagram followings but this type of online harm from misogyny to sexual abuse has resonated with women across the globe. All women deserve to be safe online, and products and services should be designed with safety in mind, including the need to have effective reporting mechanisms for the abuse and dealing with serial and coordinated abusers.

Ms Hartshorn-Sanders, you mentioned, I think I want to make sure I heard correctly, you believe that you have evidence that Instagram is still even today, failing to take down 90% of inappropriate content that is flagged to it.

100. Yes, that is correct. That was one of the findings in our misogyny report, *Hidden Hate* which also correlates with our recent findings on Anti-Muslim hate where they failed to take down 84% of anti-Muslim and Islamaphobic content.

101. Failing to act on online hate and misinformation is not the only problem that Instagram have (see, for example, our report *Malgorithm*, which revealed the problems with their algorithm driving users towards anti-vax content and conspiracy theories).

First, the ability of designated user representation groups to raise super complaints and the kind of issue you just mentioned, a systemic issue could be the subject of such a super complaint to Ofcom. In this case, it will be about Instagram. And secondly, clause 18 of the bill imposes duties on the platforms to have proper complaints procedures, where they've got to actually deal with these kinds of complaints properly. Do you think those two provisions in the bill with super complaints mechanisms for representative groups and the complaints procedure in clause 18 would go a long way towards addressing the issue that you've very helpfully quite rightly identified?

102. Under the safety duties protecting adults, Category 1 providers need to have provisions in their Terms and Conditions about how they will deal with lawful but harmful content (cl.13(4)), and adults need to then opt in to user empowerment duties in order to get the benefit of reducing the likelihood of seeing that content/alerts to the user about that content (cl. 14).

103. Clause 17(2) places a positive duty on Category 1 services to operate "systems and processes that allow users to easily report content, which they consider to be content specified" - i.e. includes content that is harmful to adults. This sits alongside the clause 18 duty to operate a complaints procedure that:
    a. "allows for different types of complaints to be made", which includes (of relevance to this):
        - Clause 17 (content reporting) - applies to all services; and
        - Clause 13 (adults' online safety duties) - applies to Category 1 services;
    b. "Provides for appropriate action to be taken by the provider of the service in response to complaints of a relevant kind"; and
    c. Is easy to access, easy to use (including by children) and transparent.

104. This is supported by a supplementary duty in clause 18(3) where all regulated user-to-user services are required to include in their terms of service provisions which are easily accessible (including to children) specifying the policies and processes that govern the handling and resolution of complaints of a relevant kind.

105. Ofcom's audit notice and power is only in relation to enforceable requirements - Schedule 11. This includes the duties outlined above, which means that they will be scrutinising at the system level, though presumably individual people could make a complaint to them that one of the platforms was failing to fulfil their duties because of their individual situation - which could trigger OfCom to decide to make an audit or take enforcement action (in accordance with their enforcement guidelines).

106. The super complaints proceedings are only available to eligible entities in regulations made by the Secretary of State. It is possible that this would extend to organisations like CCDH who could raise the failure to act and problems identified through, for example, our research on problems with Instagram's algorithms, though that detail remains to be seen. Super complaints only apply in respect of regulated services - i.e. those in scope of the Bill.

107. Abuse over messenger-only applications, which seems to include, for example, WhatsApp and Signal, appear to be excluded from the scope of the Bill under the definition of "user-to-user" service by virtue of the exemption in Schedule 1. Given that they are owned by the same parent company - i.e. Meta - Instagram may follow Facebook's lead of separating out the messenger service from their main app, though note that they have this as an integrated part of the service on the website version. It would be good to get clarity from the drafters as to whether these "separate" but connected features of the service are outside the scope of the legislation or not. Abuse over messenger services and the failure of providers to act on this type of abuse were the subject of our analysis in Hidden Hate - this is a known problem.

Do you think there should be more in the bill around a specific reference to violence against women and girls abuse and threats? And misogyny?

108. There is obviously a major issue with harmful content targeting women and girls both in the public sphere through misogynist and violent extremist content, small sites dedicated to online hate, see for example, our work on Incel sites, and in the private sphere, as evidenced by our recent report, Hidden Hate. Instagram's failure to deal with 9 in 10 reports of misogyny is largely similar to the failure of all platforms to act on other forms on online such as (see our research on anti-black racism, anti-semitism and anti-Muslim hate).

109. There are benefits that will arise from the major system changes that are created by the Bill for all these groups. The recommendations that we have made about small sites and the inclusion of misinformation / disinformation on the face of the Bill will also help target these problem areas. Further thought should be given to the intersectionality of this abuse, and the way that this impacts on women when OfCom designs their Code of Practice and audits compliance.

110. Misogyny that falls short of an offence should be included in the lawful but harmful category for the purposes of the Bill.

APPENDIX A: Summary of Recent Research

ANTI-MUSLIM HATE

CCDH researchers identified and reported 23 groups dedicated to anti-Muslim hatred and 530 posts with 25 million views to the platforms. Facebook, Instagram, TikTok, Twitter, and YouTube **collectively took no action on 89% of posts containing anti-Muslim hatred and Islamophobia.**

- Tech platforms **failed to address 89% of posts promoting the "Great Replacement"** conspiracy theory — violating pledges made following the 2019 Christchurch mosque terror attacks and signing on to the Christchurch Call.

- Facebook **failed to take action against 94% of posts promoting anti-Muslim hate**; Twitter 97%; YouTube 100%; Instagram 86%; and TikTok 64% — even after this content was flagged to moderators.

- Facebook hosts several groups dedicated to spreading anti-Muslim hatred, with a combined 361,922 followers.

- These findings echo CCDH's previous 'Failure to Act' reports. Earlier last month, researchers found that Instagram fails to act on 90% of user reports of misogynist abuse sent via Direct Message, and in 2021 CCDH discovered that Big Tech platforms collectively ignore 84% of antisemitic posts.

- While the study cited above focused on hate towards Muslims, the 'Great Replacement' conspiracy theory was one of the key themes of hate content that CCDH found and analyzed. The 'Great Replacement' conspiracy has been weaponized against many minority communities, including the recent terrorist attack in Buffalo.

MISOGYNY (Hidden Hate): CCDH worked with five women[5] with large Instagram followings (a total of 4.8 million followers) to investigate whether Instagram (Meta) was fulfilling its public promise of putting people first and actually enforcing their policies to prohibit hate speech, including misogyny, homophobia and racism, nudity or sexual activity, graphic violence, threats of violence.

We analysed 8,717 direct messages ("DMs") sent to participants[6], which showed that:

- 1 in 15 DMs break Instagram's rules on abuse and harassment. Researchers recorded 125 examples of image-based sexual abuse (IBSA), which the UK government recently announced would become illegal.

[5] Amber Heard, actress and UN Human Rights Champion; Rachel Riley, broadcaster and CCDH Ambassador; Jamie Klingler, co-founder of Reclaim These Streets; Bryony Gordon, award-winning journalist, and mental health campaigner; Sharan Dhaliwal, founder of South Asian culture magazine Burnt Roti.

[6] This analysis is based on Instagram 'data downloads' sent by Rachel Riley, Jamie Klingler, and Sharan Dhaliwal. Amber Heard and Bryony Gordon were not able to obtain full data downloads.

- 1 in 7 voice notes sent to women were abusive, and Instagram allows strangers to place voice calls to women they don't know.
- Instagram failed to act on 90% of abusive DMs reported using the platform's tools.
- 1 in 15 DMs sent by strangers to high profile women violate Instagram's Community Standards.
- Instagram failed to act on 9 in 10 violent threats sent to women over DM and reported using its tools.
- Instagram failed to act on any of image-based sexual abuse reported using its tools within 48 hours.

The fact is Instagram is failing to enforce their own terms and conditions 90% of the time. This failure to act is consistent with our previous research that shows on receiving user reports, platforms are failing to act on:

- 87.5% of Covid and vaccine misinformation
- 84% of content featuring anti-Jewish hate
- 94% of users sending racist abuse to sportspeople
- Users who repeatedly send hateful abuse.

From this study, researchers identified a need for robust legislation and several systematic problems that Instagram must fix:

- Users cannot report abusive voice notes that accounts have sent via DM
- Users must acknowledge "vanish mode" messages to report them
- Instagram does not automatically consider previous abusive messages
- Instagram's "hidden words" feature is ineffective at hiding abuse
- Users can face difficulties downloading evidence of abusive messages

RUSSIAN PROPAGANDA (Ukraine War): CCDH research shows that Facebook has once again come short on its promises to enforce its own rules by failing to label 91% of posts containing Russian propaganda about Ukraine. Facebook announced in October 2019 that it would start "labeling state-controlled media on their Page and in our Ad Library." Examples of articles that Facebook is failing to label in posts include claims that:

- Ukraine planned a "false flag" incident "prepared by British-trained saboteurs" (RT.com)
- "American mercenaries" "preparing a provocation using chemical weapons" (RT.com)
- UK intelligence reports about an invasion are "false stories" (RT.com)
- Media reports about troop movements are "hysteria" (RT.com)
- "War-hungry armed Americans in combat clothing" are operating in Ukraine (RT.com).

The most popular article in the sample amplified comments from the Croatian President, Zoran Milanovic, that "blamed the US for escalating the [Ukraine] crisis", receiving 9,416 likes, shares and comments in total. Crucially, this is a 24 hour study

that was limited to English-language posts, which has domestic implications for sewing distrust and disinformation into English-language democracies.

METAVERSE: CCDH researchers spent 11 hours on VR Chat—the most reviewed social app in Meta's (formerly Facebook) VR Metaverse—and found that it was rife with abuse, harassment, racism and pornographic content. In fact, on average our researchers reported disturbing behavior every seven minutes such as:

- Minors being exposed to graphic sexual content
- Bullying, sexual harassment and abuse of other users, including minors
- Minors being groomed to repeat racist slurs and extremist talking points
- Threats of violence and content mocking the 9/11 terror attacks.

CCDH reported all of the disturbing incidents to Meta using their web reporting tool. All of CCDH's reports about users who abused and harassed other users went unanswered. This research was recently featured by NBC News and the New York Times in its own reporting on the dangers of the Metaverse.

CLIMATE DISINFORMATION: The Toxic Ten, ten publishers who spread baseless, unscientific climate denial on their own websites and across social media, are responsible for 69% of all interactions with climate denial content on Facebook and collectively have 186 million followers on mainstream social media platforms. It's a climate denial propaganda machine part-funded by Google via ad revenue, and spread across the world via social media, in particular Facebook, who allow them to pay to promote their denial. A new audit performed in March revealed that Facebook is still failing to label half of posts containing Toxic Ten content, breaking its promises on labeling climate content. Meanwhile, CCDH is still identifying Google ads running on Toxic Ten articles containing climate denial, despite promises that it would stop the practice last year. The Toxic Ten generated an estimated total of $7.7 million in Google ad revenue from April 2021 to January 2022.

ANTISEMITIC HATE: CCDH's report, Failure to Protect, exposed how social media companies fail to remove user-reported content that contains many of the worst forms of antisemitism[7].

- CCDH researchers collected and reported 714 posts containing anti-Jewish hatred. Collectively, they had been viewed at least 7.3 million times. Posts were collected from Facebook, Instagram, TikTok, Twitter & YouTube between May-June.
- 84% of posts containing anti-Jewish hatred were not acted upon by social media companies. Facebook performed worst, failing to act on 89%, despite announcing new rules to tackle the problem.

[7] Instagram, TikTok and Twitter allow hashtags used for antisemitic content such as #rothschild, #fakejews and #killthejews that were used in posts identified by our report that gained over 3.3 million impressions.

- Platforms fail to act on 89% of antisemitic conspiracy theories about 9/11, the Covid pandemic and Jewish control of world affairs.
- Extremist anti-Jewish hate is not acted on: platforms failed to act on 80% of posts containing Holocaust denial, 74% of posts alleging the blood libel, 70% of racist caricatures of Jewish people and 70% of neo-Nazi posts.

ENDANGERING WOMEN + GIRLS: CCDH research found that Facebook and Google sold[8] ad space promoting so-called abortion "reversal" - a dangerous and unproven procedure. Ads were promoted to children as young as 13. This dangerous and unproven procedure is not approved by health authorities such as the FDA. A 2019 study of abortion reversal to test its effectiveness was abruptly halted when several participants experienced "dangerous hemorrhaging." Facebook's Ad Library shows that it accepted between $115,400 and $140,667 for 92 ads promoting or endorsing so-called abortion "reversal" since January 2020. According to Facebook's analytics, these ads received up to 18.4 million views, and were shown to children aged 13-17 over 700,000 times.

ANTI-VACCINE + COVID-19 MISINFORMATION

Since 2020, CCDH has studied the online architecture of the anti-vaccine movement and social platforms' involvement in the proliferation of anti-vaxx and Covid-19 misinformation. Through research of the tactics, organizations, revenue streams, and use of social media by leading anti-vaxxers, CCDH has established a portfolio of research on the subject.

MALGORITHM: Instagram's new algorithm, launched August 2020 amidst the pandemic, publishes recommended posts containing vaccine misinformation to users not following anti-vaxx accounts. We discovered the algorithm also cross-fertilizes and converges radical world views by publishing recommended posts about hate, conspiracies including QAnon, and election lies.

- Researchers reviewed recommendations from Instagram's 'Explore' feature and the new 'Suggested Posts' feature, which publish content into users' feeds based on an algorithm that seeks to maximize user time on platform and engagement.
- More than half of posts recommended by Instagram contained misinformation related to Covid-19, followed by anti-vaxx, QAnon, antisemitism, and election misinformation.
- This follows a series of reports by CCDH on platforms' failure to act on Covid-19 misinformation, where we found that platforms failed to remove 95% of anti-vaxx and Covid-19 misinformation reported to them using their own reporting tools.

[8] Facebook continues to make money from selling this ad space for aborttion reversal content. See, for example, https://www.facebook.com/ads/library/?active_status=all&ad_type=political_and_issue_ads&country=NZ&id=354826106095796&view_all_page_id=49651563727&search_type=page&media_type=all.

DISINFORMATION DOZEN: CCDH analysis of a sample of anti-vaxx content that was shared or posted on Facebook and Twitter showed up to 65% of anti-vaccine content can be traced to twelve of the leading online anti-vaxxers, the Disinformation Dozen. The Disinformation Dozen– including Robert F. Kennedy Jr., Joseph Mercola, and Ty and Charlene Bollinger, among others– continually violate the terms of service agreements on Facebook and Twitter.

- Following CCDH campaigning, the Disinformation Dozen have lost 56 social media accounts with 7 million followers, equivalent to nearly half the followers their accounts have amassed to date.

PANDEMIC PROFITEERS: Research into the leading anti-vaxxers and their organizations revealed that anti-vaxxers represent an industry with annual revenues of at least $36 million,
based on a limited view of their finances based on self-reported filings and publicly available revenue estimates for 22 organizations belonging to twelve of the industry's biggest earners.

- The anti-vaxx industry boasts annual revenues of at least $36 million and is worth up to $1.1 billion to Big Tech with 62 million followers across their platforms.
- Anti-vaxxers have received more than $1.5 million in federal loans through the Paycheck Protection Program (PPP) designed to help businesses through the Covid pandemic.

APPENDIX B: Programmatic Advertising Transparency

What is the problem?

1. Each year, respectable companies and their customers unwittingly funnel millions of pounds directly to the Internet's most malicious and subversive actors and messages.[9] Misinformation and hate sites are almost entirely funded by online advertising—often paid for by unsuspecting mainstream organisations who don't know what content their brand is appearing next to, and thereby funding.[10]

2. The presence of mainstream, respectable organisations next to extremist content—including climate denial, anti-vaxx propaganda, political misinformation, incel ideology and racial hatred—also serves to normalise extremism.[11]

3. More information about this problem is available here: VIDEO: Stop Funding Misinformation / www.counterhate.com.

Why is this important?

4. **It impacts on companies and NGOs:** In recent years, the Center for Countering Digital Hate has found that adverts for brands such as *Chevrolet, Capital One, DHL International, Boots, Canon, Sensodyne, Paperchase, Bloomberg, the International Rescue Committee* and others have been automatically placed on sites dedicated to promoting hate and/or misinformation.

5. **It legitimises harmful disinformation and has led to dangerous offline behaviour**. For example, we know that websites promoting the following disinformation have profited from these adverts:

    - **Anti-vaccine and Covid-sceptic misinformation**

    - **Climate denial** and conspiracy theories about climate activism;

[9] https://www.cnbc.com/2017/03/20/google-online-advertising-adjacency-problem.html
[10] https://www.technologyreview.com/2021/11/20/1039076/facebook-google-disinformation-clickbait/
[11] https://www.emarketer.com/content/does-bad-content-affect-consumer-perceptions-of-brand-safety

- **Conspiracy theories promoting the "stolen election" myth** in the wake of the 2020 US presidential election, which directly led to the domestic terror attack on the Capitol on January 6 2021; and

- **Misinformation sites** promoting antisemitism, Islamophobia, misogyny and pro-Assad conspiracy theories.

How is this happening?

6. The adverts are placed by third-party brokers, such as Google's Adsense business, which then allocate adverts to particular sites in order to fulfill predetermined target demographic (age / gender / location) and psychographic (attitudinal and behavioural) profiles.

7. The use of algorithms to automatically match adverts with web pages in order to reach a target profile has led to these services being called "programmatic advertising".

8. In their zeal to maximise profit, however, brokers often agree to place ads next to harmful content. The business model does not need to consider the values of the organisations in the advertisements because there is no transparency for those organisations about where the advertisements will end up.

9. Of course, tech companies also profit from this arrangement. Google states that publishers retain 68% of the revenue generated by Google Ads on their sites, while Google retains the remaining 32%—thus providing a strong disincentive for proper oversight. We know that Google's "brand safety" systems for adverts are not fit for purpose and, at present, there is little to no transparency around online programmatic advertising.

What will the proposed amendment do?

10. The amendment will:

    - Require advertisers to publicly declare, on their websites, the domains on which their adverts appear. This creates a driver for corporate accountability that consumers' money is not being funnelled to content that fundamentally harms society.

    - This information is often provided to advertisers by brokers, some of which are updated in real time. This amendment would simply require advertisers to disclose the URLs of the pages on which their adverts appear—but not other information, such as performance data

or targeting criterion.

- It wouldn't create a duty for advertising organisations to conduct costly studies—but by making these URLs publicly available, it will make it easier for researchers, journalists, authorities and the public to instantly access the relevant information. This creates an accountability ecosystem of enabling legislation, transparent corporate behaviour and civil society/ other companies doing the checking. There are organisations such as GDI and NewsGuard that can provide the "checklist" for advertisers. CCDH's Stop Funding Misinformation has a much shorter and much more focused "Blacklist".

- It would not conflict with GDPR, because it does not involve any personal information

11. **Scope and costs**: The amendment would only apply to large organisations, thus avoiding the need for smaller entities to bear undue administrative burden. The nominal administrative costs to large-scale programmatic advertisers, such as Google, would be easily absorbed.

What impact will this have?

12. The proposed amendment is a simple, common-sense measure which would solve a social harm overnight.

13. Since CCDH launched the Stop Funding Misinformation campaign, several sites dedicated to spreading identity-based hate using misinformation have closed, after being starved of revenue derived from Google adverts. Unsurprisingly, large organisations, which invest great sums into the management of their reputations, are almost always extremely quick to pull advertising from malicious content. The advertising industry itself has been receptive to CCDH's efforts to highlight the problem and potential solutions.

14. The requirements on companies created by the amendment would nudge them towards greater corporate responsibility, knowing that others will find it easier to see if they are funding dangerous hate and misinformation sites via programmatic advertising services.

15. Transparency requirements will provide a sustained and effective corrective market measure that will have a direct impact on individual and social outcomes - from climate change to online hate. By creating a duty of care on platforms such as Google's Adsense towards its clients (large brands and

corporations) and their customers, there would be an immediate effect on arresting the unwitting financial support given to online harm actors all over the world by cutting off revenue streams for commercialised hate and misinformation.

16. In its current form, the Online Safety Bill also contains no provision for greater oversight of advertising. This would prevent OFCOM from examining the role that ads play in funding websites promoting harmful hate and misinformation.

17. An amendment to the Online Safety Bill which would require large companies to offer greater transparency over where their adverts have been placed would create a strong and instantaneous reputational incentive for firms to cut off revenue streams for commercialised hate and misinformation.

FAQs

Won't the market change this?

18. While a recent #StopHateForProfit boycott of Facebook advertising by some companies / organisations helped to draw attention to the general problem, the economic inequality and imperative to advertise online (particularly for many of the small businesses involved) meant that this was not sustainable. Many people have returned to advertising online in order to sustain their business and livelihoods, but they are still unclear to what extent, or exactly where, their advertisements are being published. There have been little systemic changes to the advertising business model or transparency.

19. There is a network of organisations which provide help to advertisers by assessing websites on which ads appear for their content . These include NewsGuard and GDI. As such, there is a viable ecosystem of commercial providers who can form part of a programmatic advertising transparency ecosystem, all of which would be underpinned by enabling legislation by HMG to create the nudge driver for accountability and responsible advertising practices.