Alfred Landecker Foundation

Pathways to Online Hate: Behavioural, Technical, Economic, Legal, Political & Ethical Analysis.

Authored by:
Prof. Mary Aiken, Sam Donaldson & Conor Tinnelly

November 2021

Perspective Economics

Contents

Authors and acknowledgements:

This report was authored by Cyberpsychologist Professor Mary Aiken (Professor of Forensic Cyberpsychology at the University of East London & Professor of Cyberpsychology at Capitol Technology University, USA); Sam Donaldson and Conor Tinnelly (Perspective Economics). Perspective Economics is an economic research firm, focusing on the impacts of emergent technologies and sectors.

The authors would like to thank the Alfred Landecker Foundation for commissioning and funding this research paper, with particular thanks to Melina Sánchez Montañés and Raphael von Aulock from the ALF Innovation Fund for their advice, support, and feedback throughout.

Further, we would like to thank the consultees that took part in the research consultations to help shape the research.

1 Introduction

The Alfred Landecker Foundation seeks to create a safer digital space for all. The work of the Foundation helps to develop research, convene stakeholders to share valuable insights, and support entities that combat online harms, specifically online hate, extremism, and disinformation.

Overall, the Foundation seeks to reduce hate and harm tangibly and measurably in the digital space by using its resources in the most impactful way. It also aims to assist in building an ecosystem that can prevent, minimise, and mitigate online harms while at the same time preserving open societies and healthy democracies.

The five pillars that inform activity undertaken by the foundation are outlined below:

- **Strengthening democracies:** invest in the digital update of democracy by strengthening democratic values and institutions and protecting them from hostility and digitally spread hatred;
- **Protecting minorities:** stand up for an open and democratic society in which minorities are protected
- **Combating antisemitism:** fight antisemitism and hatred with contemporary means and innovative project partners - especially where they are thriving: on the internet
- **Depolarising debates:** create spaces for protected encounters and for the fair exchange of ideas and views to safeguard democratic values; and
- **Confronting the past:** Awareness of the Holocaust means preserving the memory and drawing from its enlightening energy to combat antisemitism, racism, and group hatred today

With this in mind, the Foundation is well placed to tackle online hate, to support the development of a formal and coherent approach that helps define hate online, its pathways, and the varying technologies and approaches that can be used to prevent hate and support a fairer and more inclusive online landscape.

2 Methodology

A non-exhaustive literature review was undertaken to explore the main facets of harm and hate speech in the evolving online landscape and to analyse behavioural, technical, economic, legal, political and ethical drivers.. Core themes addressed in the research include:

- **Theme 1 - Behavioural Science:** social and psychological pathways to online hate;
- **Theme 2- Technological:** specific technical interventions delivering impact in both enabling and responding to online hate;
- **Theme 3 - Economic:** exploring the incentives and economic drivers underpinning the creation, facilitation, and response to online hate; and
- **Theme 4 - Legal, Political and Ethical:** regulatory, political, and ethical considerations.

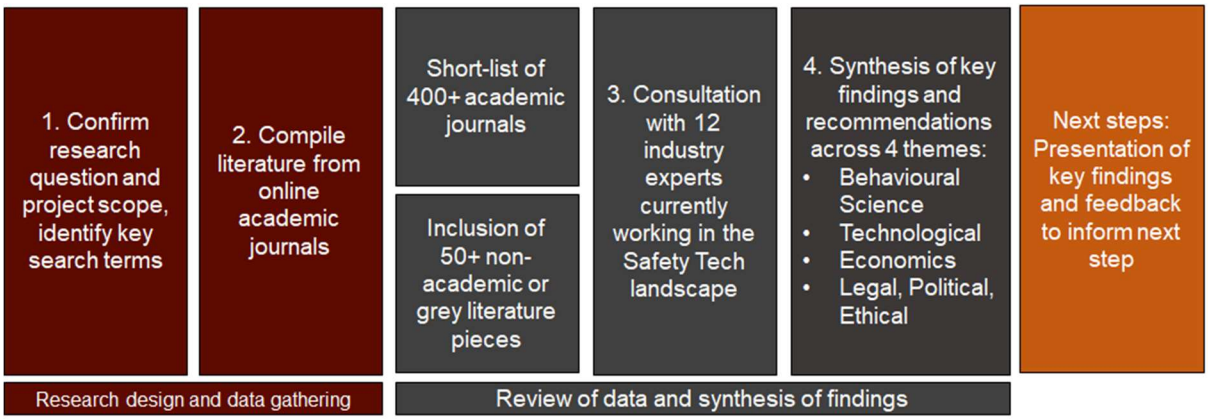
Across each of the four themes, 126 keywords were identified and then used to identify relevant papers available through Google Scholar which yielded a long list of 980 papers, which was subsequently refined to 428 papers.

Additional checks were also conducted to ensure quality and relevancy to the research area. These papers were subsequently mapped against each of the four themes and analysed and evaluated to inform key sections below.

The review also included relevant grey literature identified by the research team. The inclusion of grey literature was vital given the disparity between what can be objectively reviewed within an academic setting in a timely manner and insights that are readily available from experts in an industry setting.

A further 12 consultations with a range of expert stakeholders were also conducted and used to support key findings and recommendations outlined in this review.

Figure 2:1 Literature Review Methodology



Source: Perspective Economics

3 Theme 1: Behavioural Science

3.1 Introduction

This section provides an overview of online harm and hate speech from a behavioural perspective, offering insight into the differences and overlap in existing definitions, and how this manifests across groups.

It also looks at the different factors or "pathways" that exist both online and offline that influence the manifestation of online hate, before finally outlining the role of technology in facilitating online harm and hate speech.

> *"A reliable way of making people believe in falsehoods is frequent repetition, because familiarity is not easily distinguished from truth."*
>
> (Daniel Kahneman, 2011, Thinking, Fast and Slow)

Key findings identified from the review of *Theme 1: Behavioural Science* include:

- Existing definitions of online hate and harm are broad and often deliberately ambiguous so they can include future, unknown harmful content;
- Online hate speech differs from offline hate speech in multiple ways. It typically has a wider reach, and permanence online;
- There are a number of "pathways" to online hate, which can be linked to platform design, or pre-existing personality characteristics, and these pathways are typically multi-factorial and non-linear;
- Different platforms catalyse problematic behaviour in different ways but intervention on each can limit problematic activity and encourage problematic users to move to smaller alt-tech platforms with smaller audiences; and
- It is ineffective to see intervention as a platform problem, as hateful content and behaviour occurs across the tech ecosystem, requiring multiple interventions and education to address wider issues and limit the spread of hateful content.

3.2 Defining online harms

Online Harms Ofcom's (2019)[1] definition of online harm provides a good overview of what constitutes harm online. Outlined below this can include harmful content and hate speech, and more broadly platform design, unethical use of data, consumer rights and protecting the integrity of media. Note cyber security and fraud are included below but are often distinguished in their own right.

Table 3:1 Ofcom: Online market failures and harm definition

Category	Harm	Description
Competition policy	Competition harms	Excessive prices, limited services, lack of quality and lack of innovation, sometimes arising due to leverage market power into other markets.
Consumer protection	Fraudulent/ unfair business practices	Financial harm due to scams, distorted consumption decisions or harms to health/ wellbeing
	Unfair price Userisation	User data allowing targeted pricing - charging higher prices to vulnerable consumers, which may be considered unfair
Data protection	Harm to privacy	The nuance between targeted services (advertising), distress (distaste of surveillance), under-use of otherwise beneficial services.
	Data breach	Identity theft (e.g., financial and time cost to address fraud), cybercrime, distress, or costly actions to prevent harm following a data breach.
Cyber security	Security and resilience issues	Surveillance issues, attacks on infrastructure to undermine business or society generally.
Media policy	Risk to media plurality and quality	Too much influence over the political process by few entities, challenges to the sustainability of high-quality journalism and risks of echo chambers

[1] Ofcom. (2019). *Online market failures and harms: An economic perspective on the challenges and opportunities in regulating online services*. [online] Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0025/174634/online-market-failures-and-harms.pdf.

Content policy	Content harms	Illegal (e.g., CSAM, terrorist content) or non-illegal harmful content (e.g., age-inappropriate content, self-harm advocacy).
	Conduct harms	Cyber-bullying, trolling, intimidation, sexting (under 18s), harassment.
	Disinformation	Falsely manipulated content deliberately created or shared to deceive citizens or to cause harm for political/ user/ financial gain.
Health policy	Harm to wellbeing	Services that can lead to addictive behaviour (e.g., gambling) and excessive use (e.g., screen time)

Source: Ofcom: Online market failures and harms (2019)

In contrast to Ofcom's definition, the UK's Online Safety Bill's[2] focuses on content, conduct, and contact-based harms.

> *"User-generated content or behaviour that is illegal or could cause significant physical or psychological harm to a person. Online harms can be illegal, or they can be harmful but legal. Examples of online harms include (but are not restricted to): child sexual exploitation and abuse; terrorist use of the internet; hate crime and hate speech; harassment, cyberbullying and online abuse."* –
> Definition of online harms, UK Online Safety Bill

When looking specifically at online hate, this definition is a good starting point as it recognises the nuance between illegal and harmful content[3]. It showcases a commitment to establish a new regulatory framework that places a "duty of care" on companies to improve the design of their service or platform to improve the safety of their users online.

This "duty of care" is derived from health and safety law and places the responsibility of care on certain service providers[4] who **must moderate user-generated content** to prevent users from being exposed to illegal and harmful content online.

2 Department for Digital, Culture, Media & Sport (2021) *Understanding and reporting online harms on your online platform*, Gov.uk. Available at: https://www.gov.uk/guidance/understanding-and-reporting-online-harms-on-your-online-platform (Accessed: September 17, 2021).
3 *Online Harms White Paper: Full government response to the consultation* (2020) Gov.uk. Available at: https://www.gov.uk/government/consultations/online-harms-white-paper/outcome/online-harms-white-paper-full-government-response (Accessed: September 17, 2021).
4 Verfassungsblog. 2021. *The UK's Online Safety Bill: Safe, Harmful, Unworkable?.* [online] Available at: https://verfassungsblog.de/uk-osb/ [Accessed 18 October 2021].

In contrast to the UK which acknowledges the nuance between harmful and illegal, Germany's NetzDG *"covers only content that violates specifically listed sections of the German Criminal Code*",[5] (i.e., hate speech, advocating violence and CSAM).

Online Hate While there is no individual, globally recognised definition of online hate speech, there is some coherence in those identified, as outlined below.

> *"Any kind of **communication** in speech, writing or behaviour, that attacks or uses pejorative or **discriminatory** language with reference to a person or a group on the basis of **who they are**, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor."*
>
> - United Nations, Strategy and Plan of Action on Hate Speech[6]

> *"A **communication** on the Internet which expresses prejudice against an **identity**. It can take the form of **derogatory**, demonising, and dehumanising statements, threats, identity-based insults, pejorative terms and slurs."*
>
> - Alan Turing Institute, VSP Regulation and the broader context[7]

> *"Any **communication** that **disparages** a person or a group on the basis of **characteristics** such as race, colour, ethnicity, gender, sexual orientation, nationality, religion, or political affiliation."*
>
> - Castaño-Pulgarín, Internet, social media, and online hate speech. Systematic review[8]

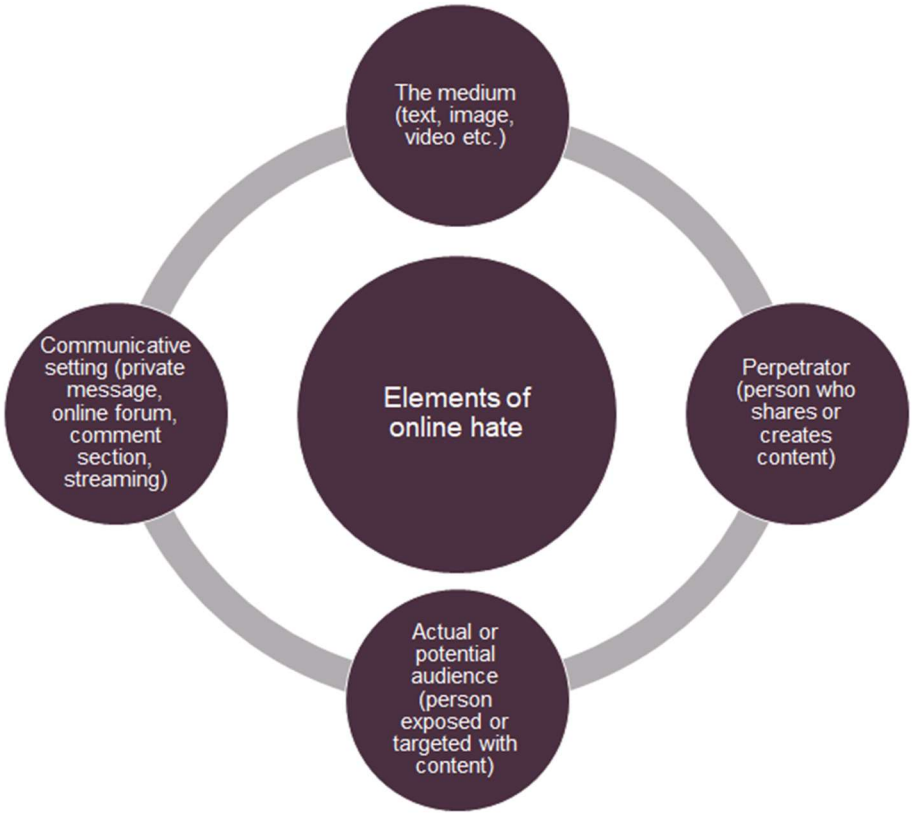
There are three core elements similar across each of the three definitions identified above, these include a communication that is in some way derogatory directed at a person based on an element of their identity.

---

5 Osborne Clarke (2020) *Online harms: The new legal framework for addressing "hate speech" in France and in Germany* Osborneclarke.com. Available at: https://www.osborneclarke.com/insights/online-harms-new-legal-framework-addressing-hate-speech-france-germany/ (Accessed: September 17, 2021).
6 Guterres, A. (2019b). *United Nations Strategy and Plan of Action on Hate Speech*. [online] Available at: https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf.
7 Margetts, H., Vidgen, B. and Burden, E. (2021) *VSP Regulation and the broader context*, Org.uk. Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0022/216490/alan-turing-institute-report-understanding-online-hate.pdf (Accessed: September 17, 2021).
8 Castaño-Pulgarín, S. A. et al. (2021) *"Internet, social media and online hate speech. Systematic review,"* Aggression and violent behavior, 58(101608), p. 101608.

The Alan Turing Foundation expands upon the above definition, outlining the different elements of online hate.[9] These are presented below and provide an alternative approach to defining online hate, one that addresses individual components as opposed to the whole process.

Figure 3:1 Elements of online hate



Source: Alan Turing Institute (2021)

The core elements include:

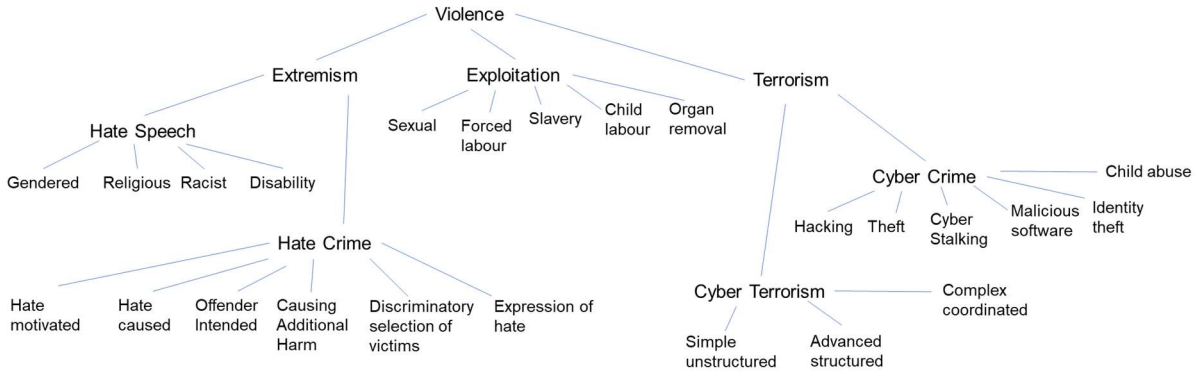
- the **medium** that the hate speech is communicated through;
- the **perpetrator** who shares the content;
- the actual or potential **audience** exposed to or targeted by the content; and

[9]Vidgen, B., Burden, E. and Margetts, H., 2021. *Understanding Online Hate VSP regulation and the broader context*. [online] Ofcom.org.uk. Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0022/216490/alan-turing-institute-report-understanding-online-hate.pdf

- the **communicative setting** (private messenger, online forum, comments, streaming, etc.)

Chetty and Alathur (2018)[10] also provide an overview of violence identified in their study on social media hate speech. The diagram below sets out the facets linked with violence and how these can manifest differently depending on its source.

Figure 3:2 Overview of violence

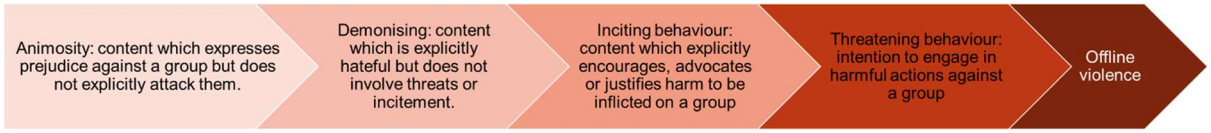


Source: Chetty and Alathur (2018)

A final consideration when defining online hate is the severity of the behaviour and the actualisation of the outcome. This is categorised below from initial animosity to demonising, to incitement, to threats and violence.

The figure shows how hateful activity can move from covert to overt hate, and potentially to offline hate or violence. Note that the development of initial feelings of animosity are not explained below but explored in the subsequent pathway sections.

Figure 3:3 Severity of online harm[11]



Source: Alan Turing Institute (2021)

[10] Chetty, N. and Alathur, S. (2018) "*Hate speech review in the context of online social networks,"* Aggression and violent behavior, 40, pp. 108–118.
11 Vidgen, B., Burden, E. and Margetts, H., (2021). *Alan Turing Institute Understanding Online Hate VSP regulation and the broader context.* [online] Ofcom.org.uk. Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0022/216490/alan-turing-institute-report-understanding-online-hate.pdf

3.2.1 The impact of online harm

Online harm can cause **immediate distress and emotional harm** initiated from viewing content, as well as l**onger-term mental health effects**, **change in victim behaviour**, **unwillingness to engage in public and civic forums**, and **incitement of hateful activity in others** (both offline and online).[12]

Those exposed to online hate speech are also likely to experience **increased levels of anxiety and depression**, **reduced levels of attachment** to their families, and **lower levels of happiness**.[13]

It can also lead to **increased tension between groups**, offline conflict, and the **normalisation of hateful activity** for majority populations.[14]

Some of the nuances that should be considered when tackling online hate include:

Permanence and reach of online hate While the online manifestation of hate may have similar presentations or impacts to offline hate[15] (e.g., online hate may generate a similar negative emotional response to offline hate) there are also some distinct features specific to online hate that should be considered. For example, online hate can be permanent[16] and easily shared across a large audience.

A symbiotic relationship between offline and online events Castano-Pulgarín et al. (2021 p. 2) state that cyberhate in general *"seems to be amplified by the use of the Internet and social networks"*. It may also be influenced by real-world trigger events, such as a terrorist attack or political hate-mongering.[17]

The symbiotic relationship between the online and offline world is noted by Slane (2007 p. 97), who argues that *"claims for the independence of cyberspace... are based on a false dichotomy... physical and virtual are not opposed; rather the virtual complicates the physical, and vice versa."*[18]

---

12 Vidgen, B., Burden, E. and Margetts, H., (2021). *Understanding Online Hate VSP regulation and the broader context*. [online] Ofcom.org.uk. Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0022/216490/alan-turing-institute-report-understanding-online-hate.pdf
13 Hawdon, James, Atte Oksanen and Pekka Ras¨anen. 2014. "*Victims of online hate groups: American youths exposure to online hate speech.*" The causes and consequences of group violence: From bullies to terrorists pp. 165–182.
14 Izsak, R. (2015) *"Hate speech and incitement to hatred against minorities in the media*." UN Humans Rights Council.
15 Danit, G. I. G., Alves, T. and Martinez, G. (2015) *Countering online hate speech*, Unesco.org. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000233231. (Accessed: September 17, 2021).
16 Crockett, M. (2016*) The internet (never) forgets*, Wpmucdn.com. Available at: https://cpb-us-w2.wpmucdn.com/smulawjournals.org/dist/8/7/files/2018/11/4_The-Internet-Never-Forgets.pdf (Accessed: September 17, 2021).
17 Awan, I. and Zempi, I. (2016) *"The affinity between online and offline anti-Muslim hate crime: Dynamics and impacts,"* Aggression and violent behavior, 27, pp. 1–8.
18 Slane, A. (2007). '*Democracy, social space and the Internet'*, University of Toronto Law Journal, 57: 81 -104

Individual vs. collective harm To understand the impact of online harm it must be understood at both the individual level and collective level. The Alan Turing Institute (2021 p. 48)[19] specifies the impact of harm on users as follows:

Table 3:2 Impact of online harm

Type of harm	Detail
Immediate negative impact on the individual	**Immediate distress and emotional harm** that individuals can experience when viewing, or being targeted by, hateful content. The harm may be heightened if the individual has been targeted by online hate previously.
Long-term negative impact on the individual	**Long-term mental health** effects of being targeted by online hate, particularly if this is combined with other forms of harmful behaviour, such as stalking and/or harassment. **Long-term impact on victims' behaviour**. Being targeted by online hate can lead individuals to change how they live their lives. In some cases, individuals report not wanting to leave their homes out of fear. **Negative effect on individuals' willingness to engage in public and civic forums**. This is possibly one of the most pernicious effects of online hate and inflicts harm at three levels: the individual, group, society
Wider negative impact on society	**Motivating and enabling offline attacks and other forms of harm**. Some hateful content directly calls on its audience to attack minority groups, whereas in other content such calls are implicit or not present and the content is best understood as 'inspiring' rather than 'inciting' harm – nonetheless, the ideas and views expressed in hateful online content may still lead parts of the audience to inflict harm on victims in an offline setting. Whether online hate leads to offline attacks depends heavily on the setting, and further research is still needed.

[19] Vidgen, B., Burden, E. and Margetts, H., (2021). *Understanding Online Hate VSP regulation and the broader context*. [online] Ofcom.org.uk. Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0022/216490/alan-turing-institute-report-understanding-online-hate.pdf

	Motivating and enabling other online attacks. Exposure to online hate can inflict other online harms: a person who views hateful content may become motivated to target individual members of a group. This could include financial attacks or scams, hacking them, doxing them (i.e., where individuals are attacked by having their private and personally identifying information shared online), or using a so-called 'cyber honeypot' to motivate the victim to engage in criminal activity (for which they could subsequently be prosecuted). Due to the sensitive nature of these other online attacks, and their resource intensiveness, this remains an under researched area. **Implications for social justice and fairness of tolerating online hate against some groups.** This is the least tangible form of hazard but is important: a society in which already marginalised and vulnerable groups are routinely harassed and attacked raises fundamental questions about its fairness.

Source: Alan Turing Institute

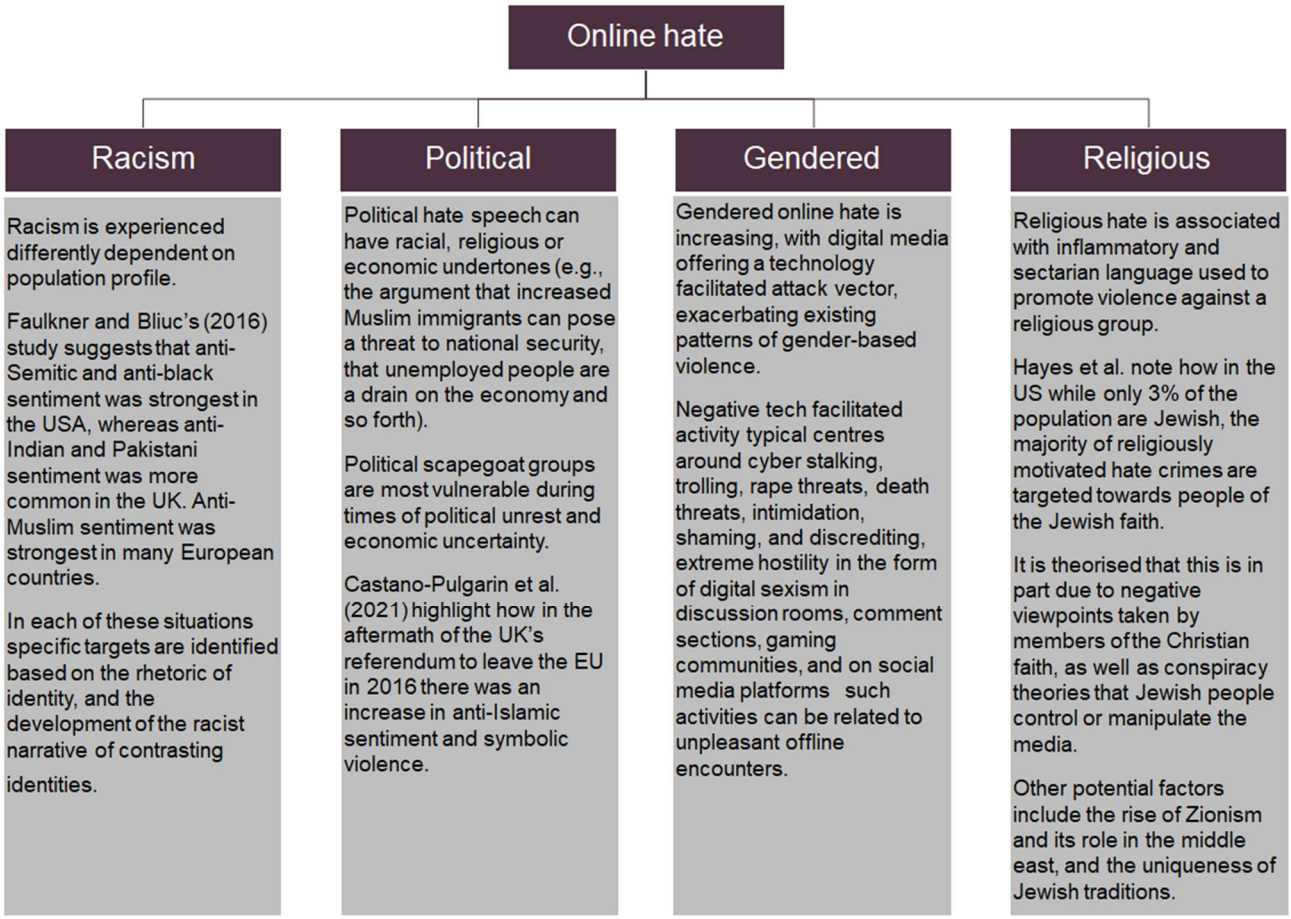
The nuance of online harm between groups Researchers argue that online hate manifestation is dependent on a range of factors such as **age, ethnicity, politics, sex, and religion**[20] and a regional understanding of these factors will be vital in combating hate speech correctly.

The Anti-Defamation League survey tracking hate speech online showcases the importance of context-driven hate. Their study revealed that online hate directed at Asian-Americans saw the most extensive single year-over-year rise compared to other groups in 2020. This is most likely related to the real-world trigger event caused by the Covid-19 pandemic,[21] which has strengthened existing prejudice in society against Asian-Americans.

20 Mossie, Z. and Wang, J.-H. (2020) *"Vulnerable community identification using hate speech detection on social media,"* Information processing & management, 57(3), p. 102087.
21 USA Today (2021) "*Exclusive: 43% of Americans say a specific organization or people to blame for COVID-19."* Available at: https://eu.usatoday.com/story/news/politics/2021/03/21/poll-1-4-americans-has-seen-asians-blamed-covid-19/4740043001/ (Accessed: September 17, 2021).

Other examples of nuance within online hate are provided in the figure below, notably splitting online hate into four broad categories (racism, political, gendered, and religious).

Figure 3:4 Nuance of online harm between groups



3.2.2 Pathways to online harms

The following section sets out literature exploring pathways to online harms. This area is arguably multi-factorial, and therefore, we first set out an overview of this literature, and consider the implications in the findings and conclusions.

The United States National Institute of Justice (2018)[22] outline factors that increase one's susceptibility to harmful behaviour online. These risk factors were identified across multiple projects and are associated both with lone actors and group acts of terrorism:

Table 3:3 Characteristics that may lead to harmful online behaviour

Both lone actor and group acts	
● Having a history of criminal violence ● Having a criminal history ● Having been involved with a gang or delinquent peers ● Being a member of an extremist group for an extended period ● Having a deep commitment to an extremist ideology ● Having psychological issues	● Being unemployed having a sporadic work history ● Lower level of education ● Lower socioeconomic status ● Failing to achieve one's aspirations ● Difficulty in romantic or platonic relationships ● Having been abused ● Being distant from one's family
Lone actor specific	
● Having a criminal record ● Having personal & political grievances ● Having received a diagnosis of schizophrenia or delusional disorder ● Having an enabler ● Being unemployed	● Having at least a bachelor's degree ● Being socially isolated ● Being single ● Living alone ● Having military experience ● Being male

Source: United States National Institute of Justice

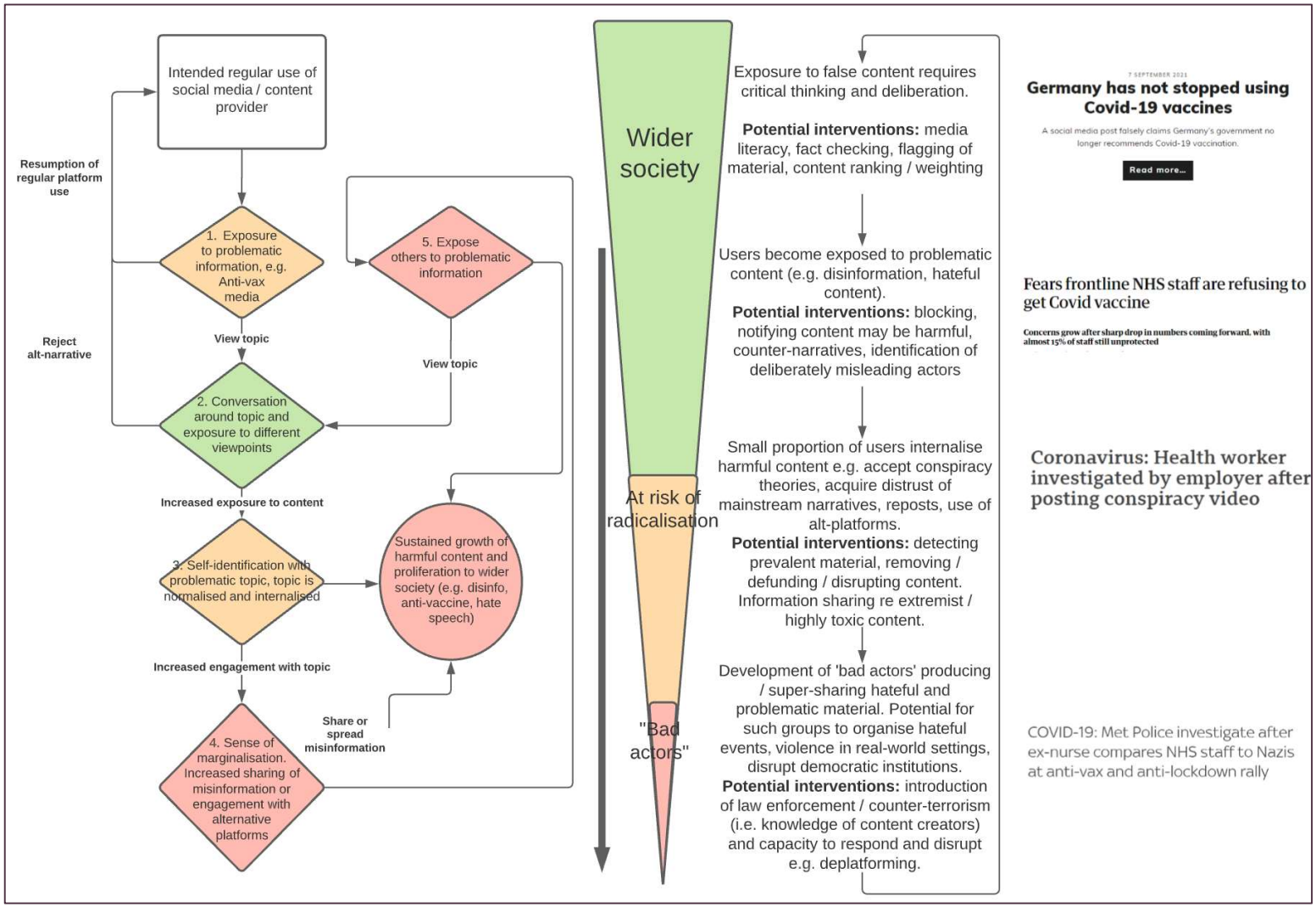
The figure below demonstrates how exposure to harmful online content can encourage deviant behaviour within susceptible individuals and the wider population.

22 Smith, A., (2021). *Risk Factors and Indicators Associated with Radicalization to Terrorism in the United States: What Research Sponsored by the National Institute of Justice Tells Us*. [online] Ojp.gov. Available at: https://www.ojp.gov/pdffiles1/nij/251789.pdf

Using COVID-19 misinformation as an example, the figure overleaf emphasises the cyclical nature of harmful content, highlighting that while exposure to problematic content is dangerous, conversation and engagement is not, if the individual rejects the alternative narrative.

When an individual begins to identify and internalise narratives this can lead to problematic behaviour, such as marginalisation from wider society, and the spread of hate speech or disinformation, or engagement in problematic real-world events.

Figure 3:5 Cyclical nature of online hate

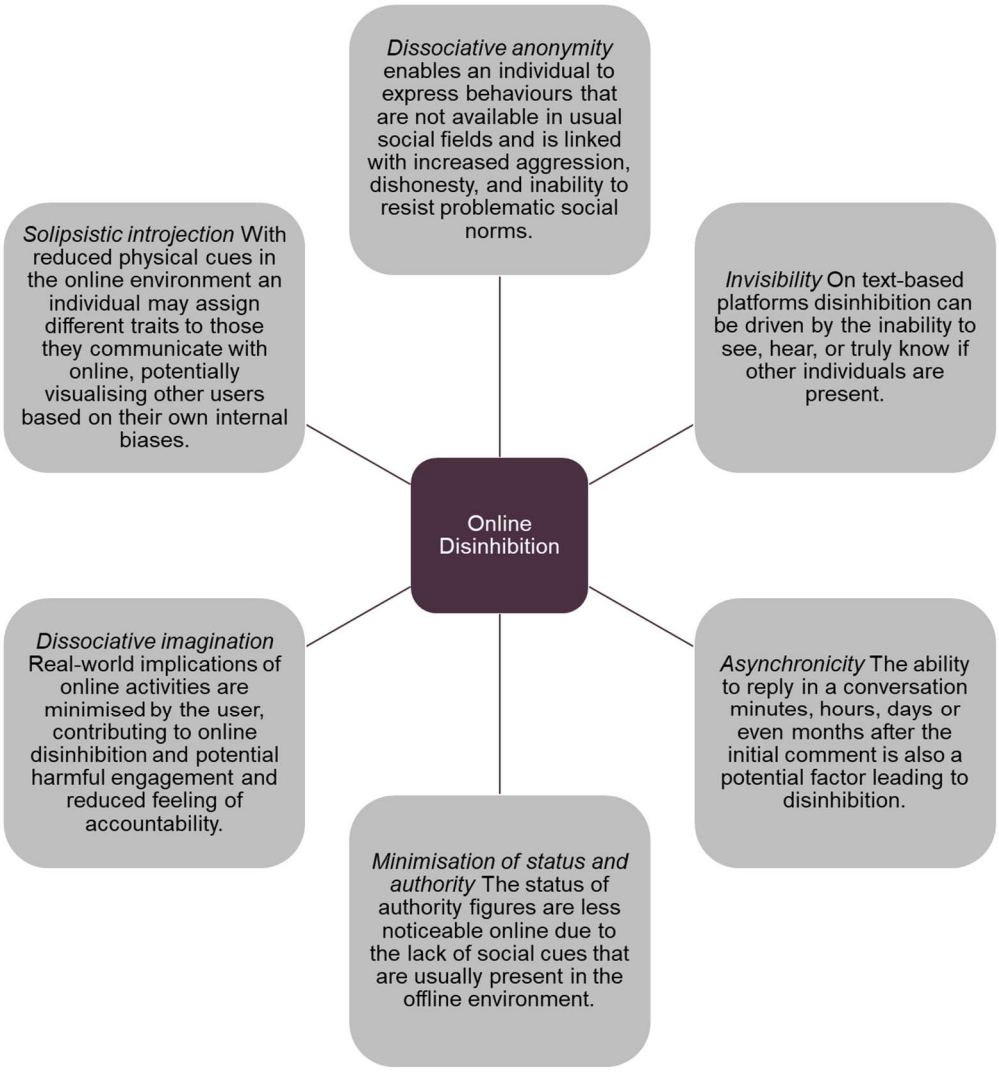


Source: Perspective Economics

Specific pathways that can catalyse the process are outlined below. These include theories grounded in traditional social psychology along with those unique to the online environment.

Online Disinhibition The Online Disinhibition Effect suggests that **inhibitions are lowered in the online environment.**[23] This can lead to a range of negative impacts such as violence, flaming and verbal attacks. Suler (2004) notes that the interaction between six psychological characteristics is responsible for most online disinhibition, each outlined briefly below:

Figure 3:6 The Disinhibition Effect



Source: Suler (2004)

[23] Wu, S., Lin, T.-C. and Shih, J.-F. (2017) *"Examining the antecedents of online disinhibition,"* Information technology & people, 30(1), pp. 189–209.

Frictionless platform design Algorithms that underlie the functionality of platforms serve to reinforce automatic behaviour, increasing impulsiveness, and potentially problematic behaviour.[24] Platform design can also lead to the formation of automatic, unconscious pattern-driven responses. This can lead to an inability to observe or judge personal internet use, classified as "*deficient self-observation*,"; or the diminished capacity to control undesired behaviour, or "*deficient self-reaction"*.[25] In both scenarios, **established automatic patterns arguably play a crucial role in losing self-control**, which is a potential antecedent for wider problematic use.

Group Polarisation Group polarisation occurs when *"members of a deliberating group move towards a more extreme point in whatever direction is indicated by the members' pre-deliberation tendency"*[26]. When platform users engage online with individuals of opposing viewpoints this can **strengthen ingroup and outgroup perspectives**.[27] This is particularly true when users are exposed to content that they strongly agree or strongly disagree with, which has a polarising effect.

A review of an alt-right group on social media platform Facebook[28] (Harel et al. 2020) provides insight into how group polarisation can occur, as well as its link to another social theory, Northrup's **Theory of Intractable Conflict**. This theory tracks the formation of in-group identity across three stages:

- **Threat**: The outgroup is perceived as a threat to ingroup identity, e.g. *'The leftists are our devil, because of their existence the country is being destroyed and the army weakened,'*;
- **Distortion:** The ingroup ceases to engage with new information relating to the outgroup, instead, distorting or dismissing information, e.g., *'I don't know if I really want to know the answer to the question of whether the thinking of the left is due to infinite stupidity or infinite naivete.*'; and
- **Rigidification:** People become locked in their position, making it difficult or impossible to change their views of the other group.

The historic design of social media applications has been known to amplify polarisation. In the past social media algorithms have been designed to elicit

24 Costa, E. and Halpern, D. (2019) *The behavioural science of online harm and manipulation, and what to do about it*, Cxmlab.com. Available at: https://www.cxmlab.com/wp-content/uploads/2019/07/BIT_The-behavioural-science-of-online-harm-and-manipulation-and-what-to-do-about-it_Single-2.pdf (Accessed: September 17, 2021).
25 LaRose, R. (2015) "*The psychology of interactive media habits,"* in The Handbook of the Psychology of Communication Technology. Chichester, UK: John Wiley & Sons, Ltd, pp. 365–383.
26 Alvernia Online. (2021). *Group Polarization in Social Psychology* | Alvernia Online. [online] Available at: https://online.alvernia.edu/articles/group-polarization-social-psychology/
27 Yardi, S. and Boyd, D. (2010) *Dynamic debates: An analysis of group polarization over Time on twitter*, Umich.edu. Available at: https://yardi.people.si.umich.edu/pubs/Yardi_DynamicDebates.pdf (Accessed: September 17, 2021).
28 Harel, T.O., Jameson, J.K. and Maoz, I. (2020b). *The Normalization of Hatred: Identity, Affective Polarization, and Dehumanization on Facebook in the Context of Intractable Political Conflict*. Social Media + Society, 6(2), p.205630512091398.

responses from its users to increase engagement, this typically includes posts that are controversial or include incendiary information.

Whilst algorithms are constantly being redesigned and updated the net result appears to be consistent, and despite main social platform Facebook's efforts to reduce political content,[29] there are still wider concerns around transparency and the efficacy of redesigns, e.g., recent whistleblower Frances Haugen[30] claim that Facebook is misleading the public, putting profit before public safety *'over and over again'*.

It is important to note that, as well as what is happening online, offline factors may also play a role in driving polarisation. Barrett, Hendrix, and Grant Sims (2021)[31] suggest that polarisation has been increasing before the advent of social media, suggesting that social media is not the main cause of rising partisan hate, and instead plays a role in intensifying divisiveness.

Echo chamber development The group polarisation effect can be strengthened if a platform's algorithm is designed to show content of interest, or that a user is likely to interact with. This can support the development of an information echo chamber.

Engagement with echo-chamber media can **confirm existing biases**, encourage users to discount similarities with their perceived outgroup, and emphasise differences (Aiken, 2018).[32]

This phenomenon has also been observed across topics such as climate change, the death penalty, affirmative action, vaccination, and sexism in male-dominated fields.[33]

A study conducted by Cinelli et al. (2021)[34] states that the "*development of homophilic (like-minded) clusters of users dominates online dynamics*". This suggests that there is homophily in the interaction networks online, as well as a **bias towards sharing information with like-minded peers**.

---

[29] The New York Times. 2021. *Facebook Dials Down the Politics for Users*. [online] Available at: https://www.nytimes.com/2021/02/10/technology/facebook-reduces-politics-feeds.html> [Accessed 18 October 2021].
[30]Pelley, S., 2021. *Whistleblower: Facebook is misleading the public on progress against hate speech, violence, misinformation*. [online] Cbsnews.com. Available at: https://www.cbsnews.com/news/facebook-whistleblower-frances-haugen-misinformation-public-60-minutes-2021-10-03/
[31] Barrett, P.M., Hendrix, J. and Grant Sims, J. (2021). *Polarization Report*. [online] NYU Stern Center for Business and Human Rights. Available at: https://bhr.stern.nyu.edu/polarization-report-page.
[32] Aiken, M., 2021. *Mass Killing and Technology: The Hidden Links*. [online] Wilson Center. Available at: https://www.wilsoncenter.org/blog-post/mass-killing-and-technology-the-hidden-links
[33]Costa, E. and Halpern, D. (2019*) The behavioural science of online harm and manipulation, and what to do about it*, Cxmlab.com. Available at: https://www.cxmlab.com/wp-content/uploads/2019/07/BIT_The-behavioural-science-of-online-harm-and-manipulation-and-what-to-do-about-it_Single-2.pdf (Accessed: September 17, 2021).
[34] Cinelli, M., de Francisci Morales, G., Galeazzi, A., Quattrociocchi, W. and Starnini, M., 2021. *The echo chamber effect on social media*. [online] Proceedings of the National Academy of the United States of America. Available at: https://www.pnas.org/content/118/9/e2023301118

Interaction with one problematic group can therefore lead to interaction with wider circles and more fringe or extreme sub-groups, normalising the problematic activity.[35] This can lead to the internalisation of group norms and increased distrust or engagement with broader society, or the out-group.

Online Syndication Online Syndication is when like-minded individuals seek out and engage with others online, fuelled by factors such as anonymity and online disinhibition, to normalise and socialise underlying tendencies.[36]

This has been seen during the COVID-19 pandemic with the rise of anti-vaccine groups. Germani and Biller-Andorno (2021) reviewed behaviour on Twitter and suggest that typically anti-vax groups are more likely to use emotive language and share conspiracy stories. The study found that the movement relies heavily on a strong **sense of community**, despite most of the content shared coming from a small number of profiles.[37] It also explains the *"Tarrant effect*" to an extent, which describes the impact of the Christchurch shooting on subsequent attacks, three out of four citing the shooter as inspiration.[38]

Marginalisation There is a growing trend of lone actors and copycat activity online, which is in part driven by a sense of marginalisation[39].

The Global Terrorism Index's report for 2020 suggests that there has been a 250% increase in far-right terrorist attacks in the West. Most of these attacks are carried out by lone actors[40] who are typically connected tangentially to or inspired by a broader ideological group or movement.

The primary issue with this is that such movements typically have a **lack of a centralised command and control structure or clear affiliation** with an organisation. The Incel movement (Involuntary Celibates) is an example of such a movement, with incidents occurring at random, with no particular method of attack.[41]

---

[35] Aiken, M., 2021. *Manipulating Fast, and Slow*. [online] Wilson Center. Available at: https://www.wilsoncenter.org/article/manipulating-fast-and-slow
[36] Aiken, M. P. (2016). *The Cyber Effect. New York*. Random House, Spiegel & Grau.
37 Germani, F. and Biller-Andorno, N. (2021). *The anti-vaccination infodemic on social media: A behavioral analysis*. PLOS ONE, [online] 16(3). Available at: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0247642.
38DE SATGE, F. (2021). *The Central Role of Memes on Alt-Right Radicalisation in the "Chanosphere."* [online] Security Destillery. Available at: https://thesecuritydistillery.org/all-articles/the-central-role-of-memes-on-alt-right-radicalisation-in-the-chanosphere
39 Torok, R. (2013) "*Developing an explanatory model for the process of online radicalisation and terrorism*," Security informatics, 2(1), p. 6.
[40] Reed, A. and Aryaeinejad, K. (2021). *2020 Trends in Terrorism: From ISIS Fragmentation to Lone-Actor Attacks*. [online] United States Institute of Peace. Available at: https://www.usip.org/publications/2021/01/2020-trends-terrorism-isis-fragmentation-lone-actor-attacks.
41Kelly, M., DiBranco, A. and DeCook, J.R. (2021). *Mass Violence and Terrorism since Santa Barbara*. New America. Available at:https://d1y8sb8igg2f8e.cloudfront.net/documents/Misogynist_Incels_and_Male_Supremacism.pdf.

Recent research to determine the scale of the Incel movement suggests that one of its biggest forums has c.13,000 active members and c.200,000 threads.[42]

It has also been suggested that anti-feminist and misogynistic communities are a gateway to wider problematic behaviours. Mamié et al. (2021) for example suggests that there is an overlap between the "Manosphere" and alt-right media.[43]

Memetic visual culture and the role of humour in radicalisation Group identity can be driven in part by **specialised visual language**. Frequently in meme format, these images can disseminate text, visual or auditory items, acting as a holder for cultural meaning for its audience, reinforcing in-group identity.

Pepe the Frog is an example of a seemingly innocuous meme that was adopted by alt-right groups in 2010 and used to legitimise politically loaded visual language of the alt-right and to validate its ideas. Meme culture has also supported the *"gamification of violence"*, with previous live-streamed terror incidents compared with the video game Call of Duty.[44]

Toxic online gaming environments The gamification of violence is a potential pathway to wider online harms in its own right and online gaming communities have in the past been targeted by terrorist recruitment teams.

As an example, Roblox has been used to recreate playable versions of infamous far-right attacks, and UK white-nationalist group Patriotic Alternative hold Call of Duty tournaments for their supporters.[45] The toxic aspect of online gaming extends across platforms.

Munn (2020)[46] suggests that video sharing platforms such as YouTube can *"create an alt-right pipeline, steering viewers from anti-SJW videos which demean so-called "social justice warriors" to gaming-related misogyny, conspiracy theories, the white supremacism of "racial realism" and thinly veiled anti-Semitism".*

---

[42] Casciani, D. and De Simone, D. (2021). *Incels: A new terror threat to the UK?* BBC News. [online] 13 Aug. Available at: https://www.bbc.co.uk/news/uk-58207064.
[43] Mamié, R., Horta Ribeiro, M. and West, R. (2021). *Are Anti-Feminist Communities Gateways to the Far Right? Evidence from Reddit and YouTube*. 13th ACM Web Science Conference 2021.
44IBID
45 Townsend, M. (2021). *How far right uses video games and tech to lure and radicalise teenage recruits*. [online] the Guardian. Available at: https://www.theguardian.com/world/2021/feb/14/how-far-right-uses-video-games-tech-lure-radicalise-teenage-recruits-white-supremacists.
46 Munn, L. (2020) *"Angry by design: toxic communication and technical architectures,"* Humanities and Social Sciences Communications, 7(1), pp. 1–11.

Amplification of existing mechanisms that support radicalisation McCauley and Moskalenko (2008)[47] determine twelve mechanisms that can lead an individual to become radicalised. These mechanisms appear on the:

- *individual level*: personal victimisation, political grievance, joining a radical group;
- *group level*: extremity shift in a like-minded group, extreme cohesion under isolation and threat, competition for the same base of support, competition with state power, and within-group competition; and
- *mass level*: jujitsu politics (mass radicalisation due to external attack), hate (prolonged violence leading to dehumanisation) and martyrdom of other group members.

While these mechanisms were not initially explored in the online context, they provide insight into potential pathways to radicalisation, which can be amplified online.

Explanations through other established social and behavioural models A range of existing psychological theories can also be used to explain susceptibility to harmful online activity, these include:

- **General Theory of Crime**, which suggests that individual factors such as self-control are directly related to crime;
- **General Aggression Model** which highlights the importance of individual and situational factors. This model emphasises the importance of the scripts and narratives developed over time which impacts how an individual reacts to an event. Example individual factors include personality, gender, beliefs, attitudes, and values; and example situational factors include aggressive cues, frustration, provocation, and incentives[48];
- **Social Ecology Theory** highlights the importance of biological elements (age and sex), history of similar behaviour, and general personality attributes (empathy, self-control); and
- **Dark Triad,** which suggests that crime can be related to one's dark personality traits such as narcissism, Machiavellianism, and psychopathy.

47 McCauley, C. and Moskalenko, S. (2008) "*Mechanisms of political radicalization: Pathways toward terrorism,"* Terrorism and political violence, 20(3), pp. 415–433.
[48] Anon, (2018). *General Aggression Model*. [online] Available at: https://www.sciencedirect.com/topics/psychology/general-aggression-model.

3.3 The role of technology in facilitating and addressing online harm

This section, while still focusing on behaviour, outlines the role of technology in facilitating harm, and the interaction between technology and an individual, and how this influences behaviour.

3.3.1 The role of technology in facilitating online harms

Technology and human behaviour have a symbiotic relationship,[49] and platform or algorithmic design can play a key role in supporting or amplifying harmful and hate behaviour, leading to the normalisation of hate speech, radicalisation, or the promotion of misleading or false narratives.

Spitaletta (2021)[50] makes a case for the existence of "Operational Cyberpsychology" which looks at mental processes in the context of interaction amongst humans and machines. This, in turn, enables "*decision-makers to more effectively understand, develop, target, and/or influence an individual, group or organisation to accomplish tactical, operational, or strategic objectives within the domain of national security or national defence"*[51]

Key examples of technology's role in influencing behaviour online are outlined below.

The normalisation of hate behaviour linked with online connectivity Increased use of social media has been linked with the normalisation of anti-Muslim hate speech and Islamophobia[52], and the Law Commission suggests that social media is now the primary medium for hate speech online.[53]

As mentioned above hateful behaviour can emerge when an individual identifies with a toxic group[54], and the dense connectivity across social networks[55] can allow hate speech and toxic ideologies to spread faster and further online.

---

[49] Licklider. (2009). *Man-Computer Symbiosis*. Available at: http://worrydream.com/refs/Licklider%20-%20Man-Computer%20Symbiosis.pdf [Accessed September 17, 2021].
[50] Spitaletta, J. A. and Hopkins, J. (2021) *Operational Cyberpsychology: Adapting a Special Operations Model for Cyber Operations*, *Nsiteam.com*. Available at: https://nsiteam.com/social/wp-content/uploads/2021/07/Invited-Perspective-Operational-Cyber-Psych_FINAL.pdf (Accessed: September 17, 2021).
[51] Staal, M. A. and Stephenson, J. A. (2013) *"Operational psychology post-9/11: A decade of evolution,"* Military psychology: the official journal of the Division of Military Psychology, American Psychological Association, 25(2), pp. 93–104.
[52] Soral, W., Liu, J. and Bilewicz, M. (2020*) "Media of contempt: Social media consumption predicts normative acceptance of anti-Muslim hate speech and Islamoprejudice."* doi: 10.4119/IJCV-3774.
[53]The Law Commission. (2014*). Hate Crime: Should the Current Offences be Extended?* [online] Available at: https://s3-eu-west-2.amazonaws.com/lawcom-prod-storage-11jsxou24uy7q/uploads/2015/03/lc348_hate_crime.pdf.
[54] Harel, T.O., Jameson, J.K. and Maoz, I. (2020). *The Normalization of Hatred: Identity, Affective Polarization, and Dehumanization on Facebook in the Context of Intractable Political Conflict. Social Media + Society*, 6(2), p.205630512091398.
55 Mathew, B. et al. (2018*) "Spread of hate speech in online social media,"* arXiv [cs.SI]. Available at: http://arxiv.org/abs/1812.01693. .

This often begins with indirect comments which are bolstered by peers and exposure to further content, leading users to use more direct and threatening language.[56]

In the online context, **anonymity** can facilitate radicalisation, enabling individuals to engage securely with people they would not engage with within a real-world setting.[57] Awan, Sutch and Carter (2018)[58] conducted corpus linguistic analysis of extremist language at different levels of anonymity and found that increased levels of anonymity were associated with an increase in extremist speech, increased conspiracy theory or media bias language, and higher levels of emotional sentiment around fear, anger, and disgust.

In another study, Mathew et al. (2019) analysed 21m posts from 341k social media users and found that *"content generated by the hateful users tend to spread faster, farther and reach a much wider audience as compared to the content generated by normal users*".[59] Further analysis found that the hateful users were *"far more densely connected among themselves*."

Costello et al. (2018) also noted how an increased engagement across social forums and the sense of identity that comes with it is linked with increased hate speech and the production of hateful material.[60]

Time spent online is, therefore, a factor in how likely an individual is to engage and identify with online groups. L1ght (2020) notes that people have spent more time on social networks, communication apps, chat rooms and gaming services during the pandemic[61] and that this can have a role in accentuating the problems endemic to these platforms, i.e., hate, abuse, toxicity, and bullying.

56 Beauchamp, Nick, Ioana Panaitiu and Spencer Piston (2018*) "Trajectories of Hate: Mapping Individual Racism and Misogyny on Twitter."* Unpublished Working Paper.
57 von Behr, I., Reding, A., Edwards, C. and Gribbon, L. (n.d.). *Radicalisation in the digital era The use of the internet in 15 cases of terrorism and extremism.* RAND Corporation.
58Awan, I., Sutch, H. and Carter, P. (2019). *Extremism Online - Analysis of extremist material on social media.* [online] Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/834369/Awan-Sutch-Carter-Extremism-Online.pdf.
59 Mathew, B., Dutt, R., Goyal, P. and Mukherjee, A. (2019). *Spread of Hate Speech in Online Social Media*. [online] Available at: https://www.researchgate.net/publication/334155686_Spread_of_Hate_Speech_in_Online_Social_Media.
60 Costello, Matthew and Hawdon.(2018) *"Who Are the Online Extremists Among Us? Sociodemographic Characteristics, Social Networking, and Online Experiences of Those Who Produce Online Hate Materials."* Violence and Gender 5(1):55–60.
61 L1GHT (2020). *Rising Levels of Hate Speech & Online Toxicity During This Time of Crisis*. [online] Available at: https://l1ght.com/Toxicity_during_coronavirus_Report-L1ght.pdf.

The role of a platform's design The online environment in which content appears is not neutral but instead planned, prototyped, and developed with particular intentions in mind.[62] Platform design can impact an array of things such as the type of participation on a platform, the type of content shown, how users interact with content, and what material is deemed inappropriate or harmful.

The table below provides an overview of seven different types of online platforms (social media, micro-blogging, video sharing platforms (VSP), gaming platforms, community forums, short-form video sites and video streaming services), highlighting how their design can influence behaviour and potentially act as a catalyst for problematic or hateful behaviour:

Table 3:4 The role of platform design in online harm

Social Media	An analysis of Facebook's design suggests the platform **promotes impulse behaviours** and that it is arguably built to be **addictive and exploitative of negative triggers**. The platform's **incentive structures and social cues for algorithm-driven media** are ideal for hate speech to develop. Facebook's newsfeed function presents information based on algorithmically generated relevance (i.e., posts receiving the most engagement) which means it is often **biased towards polarising content**. Recent research (Place, 2021) also has suggested that **'fake news' received more engagement than real news** on Facebook.[63] Facebook in the past reported during an internal meeting that by design, it was **feeding people "more and more divisive content in an effort to gain user attention** and increase time on the platform", and **prioritising controversial topics, headlines and imagery** that capture the user's attention. This is a phenomenon that has been described as **"monetising disinformation in the attention economy"** [64]

[62] Munn, L. (2020) "*Angry by design: toxic communication and technical architectures,"* Humanities and Social Sciences Communications, 7(1), pp. 1–11.
[63] Place, N. (2021) "*Fake news got more engagement than real news on Facebook in 2020, study says*," Independent, 5 September. Available at: https://www.independent.co.uk/news/world/americas/fake-news-facebook-misinformation-study-b1914650.html (Accessed: September 17, 2021).
[64] Ryan, C. D. *et al.* (2020) "*Monetizing disinformation in the attention economy: The case of genetically modified organisms (GMOs),"* European management journal, 38(1), pp. 7–18.

Micro-blogging	Twitter's primary **challenge is its trustworthiness as a news source**. The platform has been **used to covertly influence elections** in the past.[65] It has even been **linked to a change in how journalists view news sources**, with younger journalists more likely to deem tweets as newsworthy content when compared to their older colleagues. If this trend continues, the platform may **inadvertently support the emergence of pack journalism** that has broader negative influences on society's narrative. Wider issues identified within Twitter include the occurrence of **mass abuse towards trending topics**.[66] This **differs from the likes of Facebook, which would allow simultaneous threads without the volume of interactions**. An MIT study in 2018[67] found that **false news was 70% more likely to be retweeted than true stories**. Additionally, falsehoods were retweeted by unique "super sharers" more broadly than true statements. The study found that the spread of false information was essentially **not due to bots** but due to **people retweeting inaccurate** news items. [68] Online **misinformation and disinformation** have also disrupted the integrity of mainstream media. This highlights the extent to which false news stories can go undetected and have tangible consequences. A key example from 2021 is the Litecoin scandal, which resulted in disinformation - that the retailer Walmart was accepting Litecoin as legal tender - being reported by mainstream news outlets such as CNBC and Reuters.[69]
Gaming	The Anti-Defamation League[70] found that 81% of adult gamers have experienced some form of harassment online, highlighting how this occurs within their Disruption and Online Gaming framework which considers four themes when defining harmful behaviour, including: ● **Expression:** What form does the behaviour take?; ● **Delivery Channel:** Where does the behaviour occur in or around online games?;

[65] Keller, F. B. *et al.* (2020) "*Political astroturfing on twitter: How to coordinate a disinformation campaign," Political communication*, 37(2), pp. 256–280.
66 Gagliardone, I. (2015) *Countering Online Hate Speech - UNESCO*. UNESCO Publishing.
67 Felmlee, D. et al. (2020) "*Can social media anti-abuse policies work? A quasi-experimental study of online sexist and racist slurs," Socius : sociological research for a dynamic world*, 6, p. 237802312094871.
68 *Study: On Twitter, false news travels faster than true stories* (2018) Mit.edu. Available at: https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308 (Accessed: September 17, 2021).
69 Nilson, G. (2021). *Litecoin and Walmart*. [online] Finextra Research. Available at: https://www.finextra.com/blogposting/20904/litecoin-and-walmart [Accessed 18 Oct. 2021].
[70] Anti-Defamation League. (n.d.). *Disruption and Harms in Online Gaming Framework*. [online] Available at: https://www.adl.org/fpa-adl-games-framework#introduction [Accessed 18 Oct. 2021].

- **Impact:** What is the impact? (who is affected, in what ways, and what are the consequences?); and
- **Root cause:** Why does it happen? What is it an expression of?

Their framework typically finds that conflict emerges through:

- **Unintended disruptions:** These include miscommunications or unknown disruptions;
- **Aggravation:** Activity that pesters, bothers, annoys, causes grief or inhibits another player's experience;
- **Antisocial actions:** Other overly antagonistic, alienating attitudes can manifest in the context of the game;
- **Cheating:** Exploiting the rules of the game to gain an advantage or to disrupt play;
- **Harassment**: Seeking to intimidate, coerce, or oppress another player in or outside of a game;
- **Hate:** Verbal or other abuse based on a player's perceived identity;
- **Extremism:** A religious, social or political belief system that exists substantially outside of belief systems more broadly accepted in society;
- **Dangerous speech:** Content that increases the risk that its audience will condone or participate in violence against members of another group;
- **Inappropriate sharing:** Any sharing of information or content that is uninvited; and
- **Criminal or predatory conduct:** Conduct that should be escalated to law enforcement and have criminal repercussions.

Their framework also outlines where disruption and harm take place, including **in-game communication** (text chat, voice chat and emotes); **in-game mechanics** (body blocking, sabotage, cheating or manipulation, and in-game objects and imager); within the **meta-game systems** (bots, malicious reporting, advertising and monetisation of inappropriate technology or exploits, stream sniping, guild-on-guild harassment, inappropriate avatar names, harassment via the "friend" ecosystem); and **within the broader ecosystem** (social media, live streaming, harassment out of game).

Ultimately the ADL suggests that an increased attempt to moderate and control behaviour is not a sustainable path, and that anti-hate measures need to be incorporated alongside privacy-by-design.

Community forum	The most popular community forum globally is Reddit, hosting over a million individual communities. Once classifying itself as the front page of the internet, the site now refers to itself as a place for "open and honest" conversation.[71] This conversation spans an expanse of topics, legitimate news, and **wider conspiracies** (such as r/pizzagate which was home to 20k subscribers). Users on the site are **anonymous**, meaning they can explore wider interests without fear of real-world ramifications. Having said this, the open conversation on the site has in the past had **real-world impacts**, such as the witch hunt for the Boston Bomber which led to unwarranted abuse towards an innocent person.[72]
Video-sharing platform	Users are likely to **spend more time browsing and viewing content** on YouTube than other networks. Length of time on the network supports **a more subtle shift in ideology** over time through videos recommended on the homepage and sidebar, which account for 70% of all videos watched on the platform. Like Facebook, **more divisive videos attract user attention**. This informs the system's **AI to recommend similar videos** to the user and elsewhere in the network. It also **encourages the content creator to produce similar content**, arguably **creating a feedback loop for content viewed and generated and the acclimatisation of users to hateful ideology and beliefs.** In a recent paper analysing approximately 330,925 videos across 349 channels, a study found that "*users consistently migrate from milder to more extreme content"*, shifting from viewing so-called Alt-Lite material to more strident Alt-Right channels.[73]

[71] Marantz, A. (2019). *Reddit and the Struggle to Detoxify the Internet*. [online] The New Yorker. Available at: https://www.newyorker.com/magazine/2018/03/19/reddit-and-the-struggle-to-detoxify-the-internet.
[72] Reddit apologises for online Boston "witch hunt." (2013). BBC News. [online] 23 Apr. Available at: https://www.bbc.co.uk/news/technology-22263020.
[73] Munn, L. (2020). *Angry by design: toxic communication and technical architectures*. Humanities and Social Sciences Communications, [online] 7(1), pp.1–11. Available at: https://www.nature.com/articles/s41599-020-00550-7.

Short-form video

TikTok content is made up of **short (less than 60 seconds) videos** and presents an endless stream of shorts based on your **algorithmic preferences**. Within three years it has amassed 3bn users and while the network is now building out localised teams there are some major concerns around the platform.

It was reported in 2019[74] that the regulation on the platform was on a similar level to other existing social networks five years prior. During this period there were also wider **concerns about how the platform moderates content**, with the platform **hiding videos from people with disabilities, from ethnic minorities, or from LGBT users**.

Before building out regional teams globally the moderating team was based centrally in China. This has led to concerns around political censorship and censorship due to a lack of understanding of cultural nuance.

L1ght's algorithm found several worrying trends on TikTok. Their concerns were also echoed by the National Centre on Sexual Exploitation who have noted how **content on the platform is often hyper-sexualised** and that there were instances of grooming.

The growth of popularity in TikTok has raised several online safety concerns that have not existed in previous platforms. The platform's **preference for global or trending content** (as opposed to friend-based content) makes it **harder to determine harmful networks** and may stop abusive or extremist content from being noticed.

The video medium, which has already been replicated by Snapchat and Instagram, also means it is harder to pre-moderate content using a keyword analysis. Similar to YouTube the platform may also present a "rabbit hole" of **extremist or disturbing content, with the algorithm preferencing similar content** to what the user previously viewed.

74 Murphy, H. and Yang, Y. (2019) "*TikTok rushes to build moderation teams as concerns rise over content*," Irish times, 20 December. Available at: https://www.irishtimes.com/business/technology/tiktok-rushes-to-build-moderation-teams-as-concerns-rise-over-content-1.4121460 (Accessed: September 17, 2021).

Streaming	Twitch is a platform designed for streaming (primarily gaming, but other categories exist such as Music & Performing Arts, Sports & Fitness, ASMR, and Just Chatting). While Twitch **does not allow streamers to use threatening language or show sexually explicit content,** they do allow for it to be shown if it is included in the streamed content but not the main focus. Despite this, the platform's live stream function has been **used to broadcast hate crimes** in the past, such as an attack on a German synagogue.[75] In this scenario, Twitch removed the video after it was watched by 5 live viewers and 2,200 other viewers. As with the Christchurch shooting the video's hash was then shared with industry peers and law enforcement. CNBC however report that the video was shared to the Telegram network before removal and that through their own investigation they were able to find additional copies of the video on sites such as 4Chan[76].

Source: Perspective Economics

Network flow of hateful content from fringe communities to the mainstream Fringe communities on websites such as 4chan and Reddit have had a reported impact on mainstream platforms. It is therefore important to consider how harmful content is addressed both on individual sites and as part of a wider network.

While mainstream platforms may moderate or filter hateful topics, platform users are free to post and share links and opinions from other sites. Zennettou et al. (2017)[77] suggest that these links are often the source for hateful content or disinformation, which can bypass and reduce the overall effectiveness of moderation activity.

Currently, sites like Twitter use a range of solutions that prevent the spread of malicious URL content. Having said this there is still a risk of harmful content emerging on platforms. The Knight Foundation (2019)[78], as an example, identified clusters of accounts on Twitter that were linked to more than 600 fake and conspiracy news sites. The foundation found that 79% of the tweets that link to fake and conspiracy news were concentrated on just 24 outlets and while Twitter has claimed to have cracked down on automated accounts that spread this type of information, 83% of mapped accounts were still active in 2021.

[75] McGuire, K. (2020) *The Shady Side of Twitch*, *Looper.com*. Looper. Available at: https://www.looper.com/263700/the-shady-side-of-twitch/ (Accessed: September 17, 2021).
[76] Haselton, T. and Graham, M. (2019) *About 2,200 people watched the German synagogue shooting on Amazon's Twitch*, *CNBC*. Available at: https://www.cnbc.com/2019/10/09/the-german-synagogue-shooting-was-streamed-on-twitch.html (Accessed: September 17, 2021).
77 Zannettou, S. et al. (2017) "*The Web centipede: Understanding how Web communities influence each other through the lens of mainstream and alternative news sources,"* arXiv [cs.SI]. Available at: http://arxiv.org/abs/1705.06947.
78 The Knights Foundation (2019). *Disinformation, "fake news" and influence Campaigns on Twitter*. [online] Knightfoundation.org. Available at: https://knightfoundation.org/features/misinfo/.

De-platforming hateful users Established providers are known to ban specific accounts from their platforms, e.g., President Trump's Twitter account.[79] This can lead followers to migrate to platforms with fewer restrictions.[80]

Ali et al. (2021)[81] assess the effects of de-platforming users, looking at user migration from Twitter and Reddit to Gab. Blocking users led to increased activity on new platforms, but a lower reach. This suggests that banning users on one platform will prevent exposure for more moderate users, but may ultimately direct users to a smaller, but more concentrated and potentially extreme network. Ali et al.'s study notes how the process of de-platforming is not well understood currently but notes that violent actors used platforms like Gab before their attacks, e.g. Robert Bowers' anti-Semitic posting on Gab just before murdering eleven and wounding six people at the Tree of Life synagogue in Pittsburgh.

An analysis of content across platforms conducted by Kor-Sins (2021)[82] suggests that this migration is partly due to how acceptable alt-right opinions are across mainstream platforms. The article finds that Twitter's focus on politics and civil conversation is inhospitable to alt-right content. Reddit's somewhat neutral positioning and decentralised moderation system make alt-right content possible but unpopular. Finally, Gab provides a haven for alt-right beliefs, constructing its platform around "free speech" and alt-right extremism.

In a separate study that looks at the de-platforming of 3 high profile accounts (namely Alex Jones, Milo Yiannopoulos, and Owen Benjamin), Jhaver et al. (2021)[83] found that conversation mentioning these individuals dropped significantly on mainstream platforms, and that de-platforming was more successful if undertaken across multiple platforms.

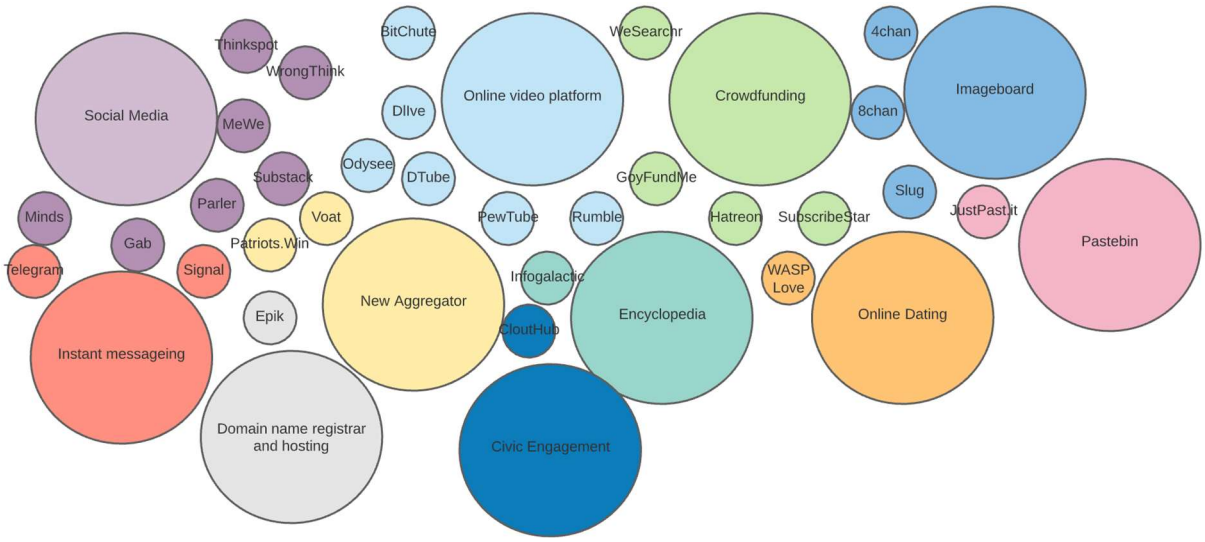
> *"I spent years growing and developing and investing in my fans …I can't make a career out of a handful… There's no future to Telegram for social media refugees if this is the best it gets... I'll just retire from social media entirely tbh… It's pathetic. So demoralising. I'm not going to waste myself on an audience of 2,000. I just refuse."*
>
> Milo Yiannopoulous following his ban from mainstream platforms

79 Twitter (2020) *Permanent suspension of @realDonaldTrump* .Twitter.com. Available at: https://blog.twitter.com/en_us/topics/company/2020/suspension (Accessed: September 17, 2021).
80 Deutsche Welle (www. dw.com) (2020) *US: Trump fans choose Parler over Twitter*, Www.dw.com. Deutsche Welle (www.dw.com). Available at: https://www.dw.com/en/donald-trump-twitter-parler-free-speech/a-55582802 (Accessed: September 17, 2021).
81 Ali, S., Saeed, M.H., Aldreabi, E., Blackburn, J., De Cristofaro, E., Zannettou, S. and Stringhini, G. (2021). *Understanding the Effect of Deplatforming on Social Networks*. 13th ACM Web Science Conference 2021. [online] Available at: https://seclab.bu.edu/people/gianluca/papers/deplatforming-websci2021.pdf.
82 Kor-Sins, R. (2021). *The alt-right digital migration: A heterogeneous engineering approach to social media platform branding*. New Media & Society, p.146144482110388.
83 Jhaver, S., Boylston, C., Yang, D. and Bruckman, A. (2021). *Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter*. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), pp.1–30.

Having said this, the subscription newsletter platform Substack has supported a variety of de-platformed figures by allowing them to set their terms for their newsletter and their subscription costs.[84]

Wider research tracking movement across platforms is limited currently, but identified alternative platforms are outlined below and described in the appendix of this report.

Figure 3:7 Alternative Technology Platforms



Source: Perspective Economics, Wikipedia

Deepfake technology Deepfakes, misinformation, disinformation, propaganda, and post-truth media have all been associated with a wide range of online hate campaigns and strategic plans to undermine democratic society and can be designed to cause unrest or to deceive its readers.[85] Deepfake technology is essential to consider given its potential role in agenda-driven internet campaigns/ supporting awareness campaigns and threat perception is vital as the technology develops.[86]

Deepfakes have been ranked by UCL as the most worrying use of AI in terms of potential application for crime or terrorism, as the technology is challenging to detect

84 York, J.C. (2021). *The delights and the dangers of deplatforming extremists.* [online] The Conversationalist. Available at: https://conversationalist.org/2021/10/01/the-delights-and-the-dangers-of-deplatforming-extremists/ [Accessed 3 Nov. 2021].
85 Fraga-Lamas, P. and Fernández-Caramés, T. M. (2019) *"Fake news, disinformation, and deepfakes: Leveraging Distributed Ledger Technologies and blockchain to combat digital deception and counterfeit reality,"* arXiv [cs.CY]. Available at: http://arxiv.org/abs/1904.05386.
86 Giles, K. and Mustaffa, M. (2019). *The Role of Deepfakes in Malign Influence Campaigns.* [online] NATO Strategic Communications Centre of Excellence. Available at: https://stratcomcoe.org/publications/the-role-of-deepfakes-in-malign-influence-campaigns/72.

and has a wide range of possible uses such as the discrediting of public figures, disinformation campaigns, and general impersonation.[87]

A report produced by the Counter Extremism Project (2020)[88] highlights that the technology is becoming increasingly democratised, meaning it will be used more widely, highlighting how this will be an essential issue in the future as society uses social media more and more as an information source.

CEP recommends a multi-body approach to addressing the challenges offered by deepfake technology, combining technical solutions with legal and public education measures.

3.3.2 Risk and prevalence of technology-enabled harm

This section provides an overview of risks identified on social media and their prevalence.

Spread of misinformation As noted previously, fake news is 70% more likely to be shared online than traditional media. Montgomery (2020)[89] highlights the severe harm caused by this oversharing, citing a range of example situations where it leads to real-world consequences, including the 2016 US election, the ethnic cleansing of the Rohingya minority in Myanmar, and the widespread misinformation associated with COVID-19. Montgomery notes that misinformation is *"highly complex, interdependent and unstable" and can only be "mitigated, managed, or minimised, not solved".*

Abuse or harassment on social media Situations of abuse and harassment are commonly found online and are often amplified by social media algorithms due to their controversial or harmful nature.[90]

A Pew Research Centre survey[91] of US adults found that 41% of Americans had experienced some form of online harassment, with a growth in the number of people reporting severe abuse (these include physical threats, stalking, sexual harassment, and sustained harassment).

This supports the above theories around normalisation and emboldened online hate activity. The Pew Centre survey reflects a trend year-on-year, reporting more abuse

87 Greaves, M. (2020). *"Deepfakes" ranked as most serious AI crime threat*. [online] UCL News. Available at: https://www.ucl.ac.uk/news/2020/aug/deepfakes-ranked-most-serious-ai-crime-threat.
88 Farid, H. and Schindler, H.-J. (2020). *Deep Fakes on the Threat of Deep Fakes to Democracy and Society*. [online] Available at: https://www.counterextremism.com/sites/default/files/CEP-KAS_Deep%20Fakes_062920.pdf [Accessed 18 Oct. 2021].
89 Montgomery, M. (2020) *Disinformation as a wicked problem: Why we need co-regulatory frameworks*, *Brookings.edu*. Available at: https://www.brookings.edu/wp-content/uploads/2020/08/Montgomery_Disinformation-Regulation_PDF.pdf (Accessed: September 17, 2021).
90 Felmlee, D. *et al.* (2020) "Can social media anti-abuse policies work? A quasi-experimental study of online sexist and racist slurs," *Socius : sociological research for a dynamic world*, 6, p. 237802312094871.
91 Atske, S. (2021) *The state of online harassment*, *Pewresearch.org*. Available at: https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/ (Accessed: September 17, 2021).

on social media channels. 79% of those surveyed state that social media firms are doing only a fair or poor job addressing online harm or bullying.

Harmful conduct through online gaming Toxicity in online gaming has been noted as an endemic problem[92] with no simple solution due to its link with the acceptable elements of online gameplay.

As mentioned, the Anti-Defamation League revealed that 80% of all online gamers had received some form of harassment and the most common forms of harassment include offensive name-calling (80% of players), trolling or griefing (60%), personal embarrassment (55%) and discrimination (53%).[93]

The harm caused by gaming is associated with dehumanisation and online inhibition and often go unreported. Some players deem such interactions acceptable within the context. Toxic gaming behaviours are varied and include harassment, verbal abuse, and flaming.

The prevalence of toxic behaviour in online gaming has led to the establishment of the Fair Play Alliance, which encourages gamers to play fair and build harassment-free communities and the online gaming sector plays a critical role in developing content moderation tools, practices, and technologies, as the underlying game often has several vital technical components to consider.

Encouragingly, many games developers have recognised the challenges ahead with improving content moderation, and this does appear to be a market with significant potential whilst ensuring inclusivity and retention of gamers.

Further, developers' role of broader standards and expectations from community players has also been an important consideration. For example, EA's Positive Play Charter sets out consequences for players who engage in racist, sexist, homophobic or abusive behaviour.[94]

Doxing is the process by which an individual or group shares previously private information about a target, typically with a subtle or covert call to action. Generally doxing is associated with other forms of online hate and is usually motivated by gender, with women more vulnerable than men.

---

[92] Beres, N. A. *et al.* (2021) *"Don't you know that you're toxic: Normalization of toxicity in online gaming,"* in Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM.
93 Anti-Defamation League (2020) *Free to Play? Hate, Harassment and Positive Social Experiences in Online Games 2020*, Adl.org. Available at: https://www.adl.org/media/15349/download (Accessed: September 17, 2021).
[94] Electronic Arts (2020) *The Positive Play Charter*, Www.ea.com. Available at: https://www.ea.com/en-gb/news/the-positive-play-charter (Accessed: September 17, 2021).

Doxing can include revenge pornography and can result in real-world safety threats.[95] It can come in several forms, such as deanonymising doxing (revealing someone's identity online), targeting doxing (revealing information that allows someone to be reached), and delegitimising doxing (revealing information that can damage someone's reputation).[96]

Chen, Cheung, and Chan (2019) attempted to identify the prevalence of doxing behaviour in a sample of Hong Kong university students (n=2,120). Ultimately Chen et al.'s study revealed that 12% of all students surveyed have in some way taken part in doxing behaviour. These students were typically younger than other students, and more female respondents were reported to have undertaken the behaviour. This is linked with respondents' attempts to ascertain the relationship status of their doxed counterparts.[97]

While this may be considered lesser harm, more widespread examples show the extent of potential harm. Examples include Anonymous' role in exposing the contact details of 7,000 law enforcement members, the Ashley Madison scandal, which saw the blackmail of people online having affairs outside of their relationships, and the misguided witch hunt on the Reddit forum after the events of the Boston Marathon.

The risks associated with doxing play a central role in the debate around the use of government IDs to verify a user online, and the implication this may have if official documented are doxed.

[95] Eckert, S. and Metzger-Riftkin, J. (2020*) "Doxxing,"* The International Encyclopedia of Gender, Media, and Communication. Wiley, pp. 1–5. doi: 10.1002/9781119429128.iegmc009.
[96] Australia Government (no date) *Doxing trends and challenges — position statement Gov.au*. Available at: https://www.esafety.gov.au/about-us/tech-trends-and-challenges/doxing (Accessed: September 17, 2021).
[97] Chen, M., Cheung, A. S. Y. and Chan, K. L. (2019) *"Doxing: What adolescents look for and their intentions," International journal of environmental research and public health*, 16(2). doi: 10.3390/ijerph16020218.

3.3.3 Role of Safety Tech in impacting behaviour

While the above has focused on the negative role of technology in amplifying online hate this section provides a non-exhaustive overview of several interventions within Safety Tech that can lead to positive behavioural change online. **Note that an overview of the Safety Tech is outlined in further detail in Theme 2 (Technology).**

Content Moderation Chandrasekharan et al. (2017) looks at the ban of hate speech on Reddit and found that removing certain pages resulted in a drastic decrease in hate speech (c.80%) in users that remained on the site. Having said this, a larger proportion of users discontinued using the site than expected.[98]

This shows that this blanket moderation and filtering can reduce hate speech in users that stay on the site but may also steer more extreme users away from mainstream networks towards websites with fewer restrictions, e.g., Gab and Voat.

At this time there are no known applications that track problem users across platforms to facilitate de-platforming.

Friction Wu (2016) argues that a decades-long campaign to monetise attention has reached a new intensity in the age of social media where the level of engagement has become monetised.[99]

This has arguably led many platforms to design their sites to offer as frictionless an experience as possible, supporting disinhibition and passive engagement which encourages reactive as opposed to reflective engagement.[100]

The introduction of friction has worked effectively on the community forum site Reddit, where users are reminded of individual subreddit community guidelines before posting.[101] WhatsApp has also limited the number of users who could be forwarded any particular news report and labelling more clearly information that has been forwarded as opposed to authored by the sender.

[98] Chandrasekharan, E. *et al.* (2017) "*You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech,"* Proceedings of the ACM on human-computer interaction, 1(CSCW), pp. 1–22.
[99] Tarnoff, B. (2016*) "The Attention Merchants review – how the web is being debased for profit,"* The guardian, 26 December. Available at: http://www.theguardian.com/books/2016/dec/26/the-attention-merchants-tim-wu-review (Accessed: September 17, 2021).
[100] Polger, D. R. (2018) *Why we need more online friction*, *Techonomy.com*. Available at: https://techonomy.com/2018/12/need-online-friction/ (Accessed: September 17, 2021).
[101] Land, M. K. and Hamilton, R. J. (2020) *"Beyond takedown: Expanding the toolkit for responding to online hate*," SSRN Electronic Journal. doi: 10.2139/ssrn.3514234.

Other advanced versions of auto-reply for messages on the market empower those susceptible to social pressures or who lack impulse control[102]. Royen et al. (2017)[103] explore these, focusing on harassment and prompts to promote self-reflection, ultimately finding that a prompt, or a secondary delay encourages greater reflection and a reduction in harmful behaviour.

Filtering Filtering can be used to classify, demote, or exclude user-generated material from a platform, reducing user exposure to negative content that may normalise hate speech.[104]

In recent years, the use of AI techniques to identify harmful content in advance has allowed platforms such as Twitter and Instagram to introduce filters that block potentially sensitive material.

In both instances, users must click and opt-in to view the content.[105] Twitter has also added an option for users to flag their tweets as potentially sensitive. While this prevents viewers from viewing harmful content, it does not solve the problem of harm moderation but instead passes responsibility onto the user.

Third-party solutions to alter platform UI and reduce negative aspects of social media Digital self-control tools are targeted at reducing screen time or the negative aspects of social media[106] and can adjust how different platforms are viewed for the user, e.g., removing the Facebook newsfeed and promoting more targeted and social engagement (as opposed to media engagement).

Counter-speech/ Re-direction Counter-speech can mitigate or prevent radicalisation in users[107] and is best delivered as an early intervention and most successful if delivered by a credible source, e.g., Siegel and Badaan (2020)[108] note that counter-

[102]Al-Mansoori, R. S. *et al.* (2021) *Digital Wellbeing for All: Expanding Inclusivity to Embrace Diversity in Socio-Emotional Status*, *Researchgate.net*. Available at: https://www.researchgate.net/profile/Raian-Ali/publication/353526705_Digital_Wellbeing_for_All_Expanding_Inclusivity_to_Embrace_Diversity_in_Socio-Emotional_Status/links/6101c0461e95fe241a95ba2e/Digital-Wellbeing-for-All-Expanding-Inclusivity-to-Embrace-Diversity-in-Socio-Emotional-Status.pdf (Accessed: September 17, 2021).
[103] Van Royen, K. *et al.* (2017) *"'Thinking before posting?' Reducing cyber harassment on social networking sites through a reflective message,"* Computers in human behavior, 66, pp. 345–352.
[104] Policy Department for Citizens' Rights and Constitutional Affairs (2020) *The impact of algorithms for online content filtering or moderation*, *Europa.eu*. Available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/657101/IPOL_STU(2020)657101_EN.pdf (Accessed: September 17, 2021).
[105] Ullmann, S. and Tomalin, M. (2020) *"Quarantining online hate speech: technical and ethical perspectives*," Ethics and information technology, 22(1), pp. 69–80.
[106] Lyngs, U. *et al.* (2020) *"I just want to hack myself to not get distracted: Evaluating design interventions for self-control on Facebook,"* in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM.
[107] We Forum (2021) *Big tech cannot crack down on online hate alone* Weforum.org. Available at: https://www.weforum.org/agenda/2021/04/big-tech-cannot-crack-down-on-online-hate-alone/ (Accessed: September 17, 2021).
[108] Siegel, A. A. and Badaan, V. (2020) *#No2Sectarianism: Experimental Approaches to Reducing Sectarian Hate Speech Online*. Available at: https://drive.google.com/file/d/184Pov1RB3mnEnNCKbDOxotOmfVvEFD1k/view (Accessed: September 17, 2021).

speech targeting sectarian hate speech was effective if focused on the commonality of faith and supported by a community leader.

A more subtle example of counter-speech is outlined in Berry and Taylor's (2017) study, which notes that prioritising "quality" input in the comment section influences subsequent engagement and perceived norms for communication.[109]

Another example, the Redirect Method, which was developed by Moonshot, uses targeted advertising to connect people searching online for harmful content with constructive alternative messages.[110]

Inoculation Borrowing from biomedical practices, teaching an individual to spot and refute misleading claims can support the development of "mental antibodies". Inoculation messages can be delivered across three components that work in conjunction: a forewarning, a refutational pre-emption and a microdose of the misleading message (akin to introducing a small dose of the virus that is weakened, in this case, by being thoroughly refuted).

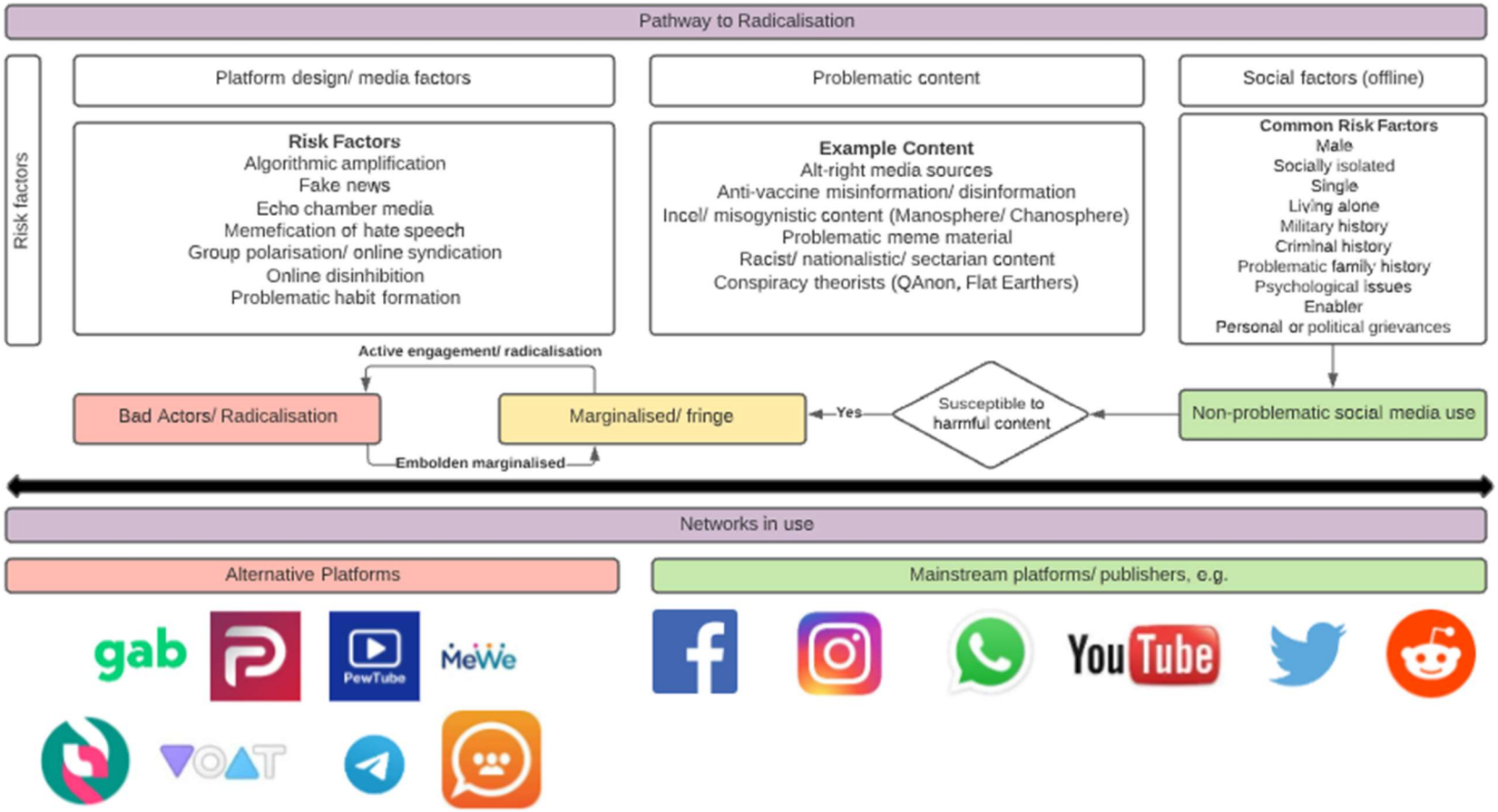
Other novel solutions Other novel solutions also exist, such as WeCounterHate's solution for Twitter which replies to hateful tweets offering a donation to rehabilitation service Life After Hate for each subsequent retweet. This solution has performed well in suppressing hate speech, on average reducing its spread by 54%, with 19% of users choosing to delete their tweet afterwards, ultimately preventing 4 million people from viewing hateful content since their launch.

[109] Buerger, C. and Wright, L. (2019) *Counterspeech: A Literature Review*, *Dangerousspeech.org*. Available at: https://dangerousspeech.org/wp-content/uploads/2019/11/Counterspeech-lit-review_complete-11.20.19-2.pdf (Accessed: September 17, 2021).
[110] Moonshot (no date) *The Redirect Method: How it works* (no date) *Moonshotteam.com*. Available at: https://moonshotteam.com/redirect-method/ (Accessed: September 17, 2021).

3.4 Summary

The figure below provides a summary of findings presented under Theme 1 – Behavioural Science.

Figure 3:8 Section summary



Source: Perspective Economics

4 Theme 2: Technological

4.1 Introduction

This section provides an overview of the Safety Tech market. This includes an overview of why there is a need for trust and safety providers within the market, the different technologies used to address harms, and how they are used across platforms. The section also outlines how Safety Tech can be used to support both preventative and proactive intervention.

> *"Technology is neither good nor bad, nor is it neutral."*
>
> (Melvin Kranzberg, 1986, first law of technology)

Key findings from the section include:

- There are at least 400 Safety Tech firms operating globally providing services that exist at the system, platform, endpoint, and information level;
- The various technologies described highlight the need for a coordinated, global approach that suppresses, marginalises, and removes harmful content. To date no single solution is known to remove harmful content in its entirety (or without wider challenges explored in this chapter);
- Safety Tech solutions can be developed in partnership with commercial partners, through academic collaboration, and with the support of public sector or industry networks and there are a range of best practice recommendations that support ethical and transparent design; and
- Trends in the market show the impact of existing measures to address online harm. These trends are also being shaped by wider factors in online activity, such as the emergence of alt-tech platforms. Evidence also suggests that greater time spent online as a result of COVID-19 among other factors has increased the risk of hate speech online.

4.2 An overview of Safety Tech

Safety Tech providers *"develop technology or solutions to facilitate safer online experiences, and protect users from harmful content, contact or conduct."*[111]

Providers in this space[112] typically develop products and approaches to:

- Work closely with law enforcement, to help trace, locate and facilitate the removal of illegal content online
- Work with social media, gaming, and content providers to identify harmful behaviour within their platforms
- Monitor, detect and share online harm threats with industry and law enforcement in real-time
- Develop trusted online platforms that are age-appropriate and provide parental reassurance for when children are online
- Verify and assure the age of users
- Actively identify and respond to instances of online harm, bullying, harassment and abuse
- Filter, block and flag harmful content at a network or device level
- Detect and disrupt false, misleading, or harmful narratives Advise and support a community of moderators to identify and remove harmful content

It should be noted that Safety Tech can also address wider issues which include brand protection and physical surveillance, as well as payment processors, data centres and AdTech, which are covered under Theme 3 of this report which outlines the economics of online hate.

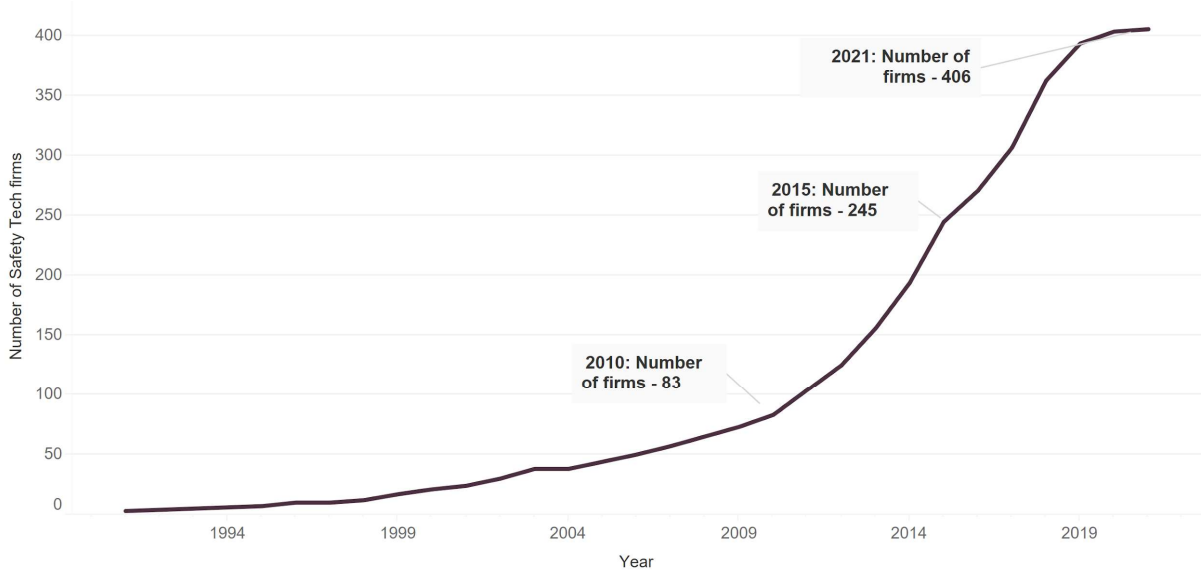
> *"It's valuable to create a name to designate a sector such as Safety Tech, it facilitates collaboration, helps to build an ecosystem and enables policy makers to recognise the sector"*
>
> (Academic stakeholder working in age verification)

[111] Department for Digital, Culture, Media & Sports (2020) *Safer technology, safer users: The UK as a world-leader in Safety Tech*, *Gov.uk*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/974414/Safer_technology__safer_users-_The_UK_as_a_world-leader_in_Safety_Tech_V2.pdf (Accessed: September 17, 2021).
[112] IBID

4.2.1 Types of Safety Tech

Ongoing work by Dealroom[113], supported by the Alfred Landecker Foundation, explores the growth of the Safety Tech market at the global level, identifying c.400 firms currently developing solutions in the sector. Growth over time is outlined in the figure below:

Figure 4:1 Number of Safety Tech firms globally



Source: Dealroom (Year = Year Founded)

The UK's Safer Technology Safer Users report (2020)[114] also classifies the various levels of use for Safety Tech technology, outlined in the table overleaf.

[113] *Global SafetyTech Ecosystem* (2021) *Dealroom.co*. Available at: https://safetytech.dealroom.co/companies.startups/landscape/f/data_type/anyof_Verified/tags/anyof_safetytech?filter=industries&sort=-revenue (Accessed: September 17, 2021).

[114] Department for Digital, Culture, Media & Sports (2020) *Safer technology, safer users: The UK as a world-leader in Safety Tech*, *Gov.uk*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/974414/Safer_technology__safer_users-_The_UK_as_a_world-leader_in_Safety_Tech_V2.pdf (Accessed: September 17, 2021).

Table 4:1 UK Safety Tech Sector Taxonomy

Technology types
System level (e.g., Microsoft's PhotoDNA)
• **Automated identification and removal of illegal content** • Use of technology to **identify and remove** known child sexual exploitation and abuse and terrorist content (especially imagery and video). *Technologies typically include hashing, URL lists, takedown and domain alerts, keyword collation.*
Platform level (e.g., content moderation providers)
• **Supporting content moderation** • **Identifying potential illegal content or conduct** (e.g., grooming, hate crime, harassment) and content that breaches the site's terms and conditions. *Technologies typically include threat detection and reporting, platform monitoring, hashing, content filtering, automated human moderation, image processing, computer vision and machine learning.* • **Enabling age-appropriate online experiences:** Use of **age-assurance** and age-verification services to limit children's exposure to harmful content. *Technologies include age assurance mechanisms, age estimates, e-IDs, database matching and attribute exchange.*
Device/ Endpoint level (e.g., SafeToNet)
• **User-initiated protection**: User, parental or device-based products that can be installed on devices to help protect the user from harm that comes in the form of software and applications • **Network filtering**: Products or services that actively filter content, through black-listing or blocking content perceived to be harmful. This includes solutions provided to schools, businesses, or homes that filter content for users that supports content filtering and monitoring
Information Environment (e.g., Snopes)
• **Identifying and mitigating disinformation:** Flagging content with false, misleading, or harmful narratives through the provision of fact-checking and disruption of disinformation (e.g., flagging trusted sources). *Technologies include disinformation research, site assurance and AI and machine learning automated fact-checking.*

Source: DCMS Safer technology, safer users (2020)

The effectiveness of individual technologies mentioned in table 4.1 above is outlined in greater detail below. These include:

Hashing Hashing supports the removal of known harmful, extremist, or damaging material that is stored on hash lists and is effectively used by a range of organisations such as the National Center for Missing and Exploited Children and the Global Internet Forum to Counter-Terrorism.[115]

It has been in use for several decades, evolving from simple randomised approaches to advanced adaptive methods that consider locality, structure, label information and data security. Hashing typically has a *data-orientated application* and a *security-orientated application*[116].

The primary issue with hashing is the lack of context afforded to images, e.g., hashing techniques have restricted the reach of posts from the Syrian Archive, a non-profit that documents war crimes where hash lists have produced false positives resulting in a takedown of information supporting the human rights of victims of war in Syria, harming the positive efforts of the Syrian Archive.

URL listing URL listing and blocking can limit access, deter, or prevent access to illegal or harmful content, e.g., in 2020, 180 URLs were identified as hosting child sexual abuse images by the Internet Watch Foundation who were able to remove 145 URLs. The remaining 35 URLs either removed their content or changed hosting countries by the time authorisation for removal was provided by the authorities. The overall number of URLs hosting such content had also increased from the previous year. This highlights three critical issues within URL blocklisting:

- that the **legal process** may be too arduous to operate effectively
- that despite blocking URLs, between 2019 and 2020, there was an increase in sites hosting indecent images (suggesting the **treatment of symptoms and not the root cause** of the behaviour), and users often have alternatives or signposts to alternatives[117] and
- that due to **geographical limitations**, indecent material may be stored overseas and beyond jurisdiction highlighting the need for a global approach.

The method is effective at introducing friction and preventing harmful content from being hosted in regions implementing the technology. The UK for example hosted 18%

---

[115]Heller, B. (2019) *Combating Terrorist-Related Content Through AI and Information Sharing*, *Annenbergpublicpolicycenter.org*. Available at: https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/05/Combating_Terrorist_Content_TWG_Heller_April_2019.pdf (Accessed: September 17, 2021).
[116] Chi, L. and Zhu, X. (2017) "*Hashing techniques: A survey and taxonomy*," ACM computing surveys, 50(1), pp. 1–36.
[117] Cory, B. Y. N. (no date) *How website blocking is curbing digital piracy without "breaking the internet," Gov.pt*. Available at: https://www.igac.gov.pt/documents/20178/557437/Estudo_2017/3adcf3b7-e9ca-497a-bebd-5fc72cec72e7 (Accessed: September 17, 2021).

of global exploitative images in 1996, down to 0.1% in 2020.[118] This does however mean that harmful content is potentially displaced rather than removed.

Takedown notice Takedown notices have been used online since the early 2000s and are often issued in response to copyright infringement, libel, or legality of content, supported by the EU's Electronic Commerce Directive (2000) and the United States' Digital Millennium Copyright Act (1998).

According to the Internet Watch Foundation, 100% of content flagged for takedown in 2019 was removed within two hours.[119] As above, the work undertaken by organisations such as the Internet Watch Foundation is focused on hosted content in the UK. Despite this, the organisation was aware of c. 8,000 additional images hosted elsewhere.

Domain alerts Domains are the first source of engagement for users online, making protection vital[120] and they can be used to report harmful or illegal material to businesses within the domain registration sector.

An example provided by the Internet Watch Foundation highlights how *"Company A"* unknowingly hosted c.560 URLs with illegal content. Following membership to IWF's domain alert service, this was reduced to 93 URLs.[121]

Another more recent example highlights how the method can be circumvented. When GoDaddy removed "prolifewhistleblower.com" for violating terms and conditions, the anti-abortion group Texas Right to Life released a statement that *"Our IT team is already in the process of transferring our assets to another provider and we'll have the site restored within 24-48 hours."*[122]

Keyword collation and monitoring Software such as Classroom Cloud[123] include 14,000 problematic terms relating to topics such as self-harm, gambling, bullying, and racism in several languages. Issues with keyword monitoring are linked with the ability to use codewords or variations of text known to flag systems, e.g., using numbers and

---

[118]Internet Watch Foundation (2020) *Face the facts The Annual Report 2020*, *www.iwf.org.uk/*. Available at: https://www.iwf.org.uk/sites/default/files/inline-files/PDF%20of%20IWF%20Annual%20Report%202020%20FINAL%20reduced%20file%20size.pdf (Accessed: September 17, 2021).

[119] Internet Watch Foundation (2019) *Annual Report 2019*, *iwf.org.uk/*. Available at: https://www.iwf.org.uk/sites/default/files/reports/2020-04/IWF_Annual_Report_2020_Low-res-Digital_AW_6mb.pdf (Accessed: September 17, 2021).

[120] Zero Fox (2019) *What is domain protection and how to address domain-based attacks* (2019) Zerofox.com. Available at: https://www.zerofox.com/blog/domain-protection-top-3-domain-based-attack-tactics-and-how-to-address-them/ (Accessed: September 17, 2021).

[121] Internet Watch Foundation (no date) *Domain Alerts*, *Org.uk*. Available at: https://www.iwf.org.uk/our-services/domain-alerts (Accessed: September 17, 2021).

[122] Reuters (2021) "*GoDaddy terminates hosting of Texas anti-abortion tip website*," 3 September. Available at: https://www.reuters.com/world/us/godaddy-terminate-hosting-texas-anti-abortion-tip-website-2021-09-03/ (Accessed: September 17, 2021).

[123] Classroom.cloud.(no date), *eSafety/Safeguarding – A helping hand,* Available at: https://classroom.cloud/online-safety/ (Accessed: September 17, 2021).

symbols to prevent word detection. Despite this, keyword collation is still used by social media firms and have even been incorporated into larger firms such as Facebook to allow group moderators to track keywords that appear in their posts.[124]

Threat detection and reporting | Platform monitoring | Artificial Intelligence AI is used to identify harms, and to support human moderation.[125]

The use of AI solutions is driven in part by a need to appease government and stakeholders and to meet the growing public expectation associated with platform responsibility. [126,127] Their use however may exacerbate wider issues across platforms linked with the opacity and ambiguity of platform governance and accountability, and the nuance of moderation.

AI bias may also complicate outstanding matters of justice (e.g., they may be biased to language, cultures, or viewpoints), or obscure or depoliticise the politics that underlie moderation.[128]

Maschmeyer et al. (2021)[129] highlight the core issue with commercial threat detection. They suggest that commercial products can prioritise high-profile victims while threats to civil society tend to be neglected or entirely bracketed. This means known threats may be overemphasised while unknown threats will be under presented.

Image processing and computer vision Sabat et al. (2019)[130] highlight how computer vision is between 2-5 years away from mainstream adoption,[131] and Forrester New Wave describes the sector as moving at "light speed". [132] Core issues identified with the use of the technology are the infrastructure requirements to operate at scale and concerns around privacy and ethical use of the technology. There are also concerns around the cost and time that goes into training a computer vision model, but this can

[124] Facebook, (no date) *Use Keyword Alerts to spot when specific terms are used in your group, Facebook.com*. Available at: https://www.facebook.com/community/whats-new/using-keyword-alerts/ (Accessed: September 17, 2021).
[125] Ehrenkranz, M. (2018) *Facebook is using new AI tools to detect child porn and catch predators*, *Gizmodo*. Available at: https://gizmodo.com/facebook-is-using-new-ai-tools-to-detect-child-porn-and-1829968486 (Accessed: September 17, 2021).
[126] Gorwa, R., Binns, R. and Katzenbach, C. (2020) "*Algorithmic content moderation: Technical and political challenges in the automation of platform governance*," Big data & society, 7(1), p. 205395171989794.
[127] Scott, M. and Kayali, L. (2020) *What happened when humans stopped managing social media content*, *POLITICO*. Available at: https://www.politico.eu/article/facebook-content-moderation-automation/ (Accessed: September 17, 2021).
[128]Gorwa, R., Binns, R. and Katzenbach, C. (2020) "*Algorithmic content moderation: Technical and political challenges in the automation of platform governance,"* Big data & society, 7(1), p. 205395171989794.
[129] Maschmeyer, L., Deibert, R.J. and Lindsay, J.R. (2020b). *A tale of two cybers - how threat reporting by cybersecurity firms systematically underrepresents threats to civil society.* Journal of Information Technology & Politics, 18(1), pp.1–20.
[130] Sabat, B. O., Ferrer, C. C. and Giro-i-Nieto, X. (2019) "*Hate speech in pixels: Detection of offensive memes towards automatic moderation*," arXiv [cs.MM]. Available at: http://arxiv.org/abs/1910.02334.
[131] Buntz, B. (2020) *2020 predictions: Computer vision projects will gain ground*, Iotworldtoday.com. Available at: https://www.iotworldtoday.com/2020/01/06/2020-predictions-computer-vision-projects-will-gain-ground/ (Accessed: September 17, 2021).
[132] Carlsson, K. (2019). *The Forrester New Wave™: Computer Vision Platforms, Q4 2019 The 11 Providers That Matter Most and How They Stack Up* [online] Available at: https://www.dlt.com/sites/default/files/resource-attachments/2020-04/The-Forrester-New-Wave%E2%84%A2_Computer-Vision-Platforms-Q4-2019.pdf [Accessed 28 Oct. 2021].

be mitigated through the commoditisation of data models that can help streamline the process.

Network identification Benigini, Joseph, and Carley (2017)[133] undertook *"Iterative Vertex Clustering and Classification,"* which is a scalable analytic approach for Online Extremist Community detection. The research team identified whether users were following or engaging with extremist groups and the direction of engagement (i.e., one way or reciprocal). They were also able to determine the relationship to the extremist group (e.g., overt supporter, affiliate, scholar), languages spoken and whether the user's account was active or suspended.

The use of similar approaches at scale has the potential to support platforms in identifying individuals at risk of engaging with toxic networks or profiles spreading or engaging prolifically within an individual network.

Content filtering The Internet Society (2017)[134] suggests that Internet blocking to address illegal content or activities is generally inefficient, often ineffective and can cause unintended damages to Internet users. It also does not remove the harmful content but instead blocks access to the site which can lead to blocking all pages on a site, not just that which contains illegal content. Wider issues include the decreased transparency and wider security risks associated with the practice, such as the requirement to view users' traffic.

Age assurance Existing procedures online are often driven by self-reported age systems that users can easily bypass by providing false information.

The age assurance market is however maturing, with the UK government's DCMS reporting 30 potential data sources for age assurance that use either officially provided, user-provided, or automatically generated information, with the greatest success rates in technologies that review official documents.[135]

A key consideration for the technology is the need to protect private information. Through the incorporation of official documents, there is the potential that sensitive information can be linked back to an official document that confirms identity, address and date of birth if leaked.[136]

---

[133] Benigni, M. C., Joseph, K. and Carley, K. M. (2017) "*Online extremism and the communities that sustain it: Detecting the ISIS supporting community on Twitter,*" *PloS one*, 12(12), p. e0181405.
[134] Internetsociety.org.(no date) *An overview of Internet content blocking* Available at: https://www.internetsociety.org/resources/doc/2017/internet-content-blocking/ (Accessed: September 17, 2021).
[135] *VoCO (Verification of Children Online) Phase 2 Report* (2020) *Gov.uk*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/934131/November_VoCO_report_V4__pdf.pdf (Accessed: September 17, 2021).
[136] Allison, P. R. (2019) *Politics, privacy and porn: the challenges of age-verification technology*, *Computerweekly.com*. ComputerWeekly.com. Available at: https://www.computerweekly.com/feature/Politics-privacy-and-porn-the-challenges-of-age-verification-technology (Accessed: September 17, 2021).

Automated fact-checking Reuters Institute[137] reveals that few people actually visit fake news sites, and those that do are usually of a particular demographic (e.g., American Republicans). Having said this, fake news stories draw disproportionate attention on social media, outperforming regular news outlets.

The potential of the technology identified by researchers and practitioners is related to its ability to assist moderators in identifying fake news, but researchers claim more must be done to improve the quality of tools.[138]

There is however a demand for the technology, with Twitter acquiring the firm Fabula AI to combat disinformation. At the time (2019) Twitter stated*: "This strategic investment in graph deep learning research, technology and talent will be a key driver as we work to help people feel safe on Twitter and help them see relevant information."*[139]

> *"There is an incredible potential for social good in the use of online safety technologies which would be most useful in tackling mis and disinformation... ML and classification techniques could be used to develop reputation and trust systems, collaborative systems and recommender systems in effect to model trust"*
>
> (Academic stakeholder working with Artificial Intelligence)

Firms offering open-source access to anonymised data Twitter engages with government and academia and is the only major service to provide a public API for research, providing a separate API for the analysis of COVID-19 specific tweets across languages.[140] The firm also ensures that it is transparent about what data it collects and how it presents tweets on its network.

[137] Reuters Institute for the Study of Journalism. (2021). *Digital News Report 2021*. [online] Available at: https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2021.
[138] Graves, L. (2018) *Understanding the promise and limits of automated fact-checking*, *Ox.ac.uk*. Available at: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2018-02/graves_factsheet_180226%20FINAL.pdf (Accessed: September 17, 2021).
[139] Twitter (2020) *Twitter acquires Fabula AI to strengthen its machine learning expertise*, Twitter.com. Available at: https://blog.twitter.com/en_us/topics/company/2019/Twitter-acquires-Fabula-AI (Accessed: September 17, 2021).
[140] Parliament (2020) *Written evidence submitted by Twitter (COR0177) Parliament.uk*. Available at: https://committees.parliament.uk/writtenevidence/5814/pdf/ (Accessed: September 17, 2021).

Additional tools Other tools that can be used to reduce harm beyond moderation have been classified by Disinfo Cloud and are outlined below:[141]

Table 4:2 Other Safety Technologies

Tool	Definition
Social Listening Tools	Assist in understanding how information is shared or spread via social media, help identify bots and trolls, or offer insight into the effectiveness of a media campaign.
AdTech Tools	Enable the targeting of content, including messaging, to relevant audiences, segmenting and targeting key audiences.
Manipulated Information Assessment Tools	These tools use contextual clues to alert users to the potential that the text, visuals, or audio they are consuming may be manipulated and may be part of disinformation campaigns.
Dark Web Monitoring Tools	Alert users to information campaigns emerging from the dark web.
Blockchain Media Authentication Tools	Technology that records a decentralised, digital record, ensuring the validity of original content and can provide a bulwark against claims of doctored media.
Fact-Checking Tools	These tools aggregate, analyse, and provide ratings on high versus low-quality information sources using a variety of metrics.
Gamified Education Tools	Games and other web-based tools that increase psychological resilience against disinformation and support critical thinking.
Internet Censorship Circumvention Tools	These tools can protect internet users against state-imposed internet censorship. They can help to facilitate the continuous flow of information and promote free expression and may also enable digital security for at-risk users.
Crowd-Sourced Content Assessments and Web Annotation Tools	These tools can help build standards of assessment of high quality versus low-quality news. Similarly, web annotation tools can enable independent dialogue and alternative viewpoints directly through online content to facilitate discussion and debate.

Source: Disinfo Cloud

[141] disinfocloud.com. (n.d.). *Disinfo Cloud.* [online] Available at: https://disinfocloud.com/tools-overview/ [Accessed 18 Oct. 2021].

4.2.2 How the major platforms use Safety Tech

Current technologies used across major platforms are presented overleaf. These include matching technologies (i.e., hashing with known content) and classification technologies (classifying newly uploaded content) and **include publicly identified moderation systems only** informed by Gorwa, Binns and Katzenbach's 2020 review[142].

Factors identified that impact what technology is used include the type of community using the platform, the type of content it must deal with, and the expectations placed upon the platform by various governance stakeholders.

Examples of the role of stakeholder pressure are presented below for two major platforms:

- Within the United States, YouTube is not legally accountable for hosting copyright-infringing material on its platform due to intermediary liability provisions in section 230 of the Communications Decency Act and section 512 of the Digital Millennium Copyright Act. Despite this, there have been growing threats to the platform, which has resulted in the use of the YouTube developed Content ID, which is designed with the copyrighter in mind.
- Twitter developed the "Quality Filter" in light of concerns around the platform's content. This filter tries to predict what content is low quality. Given Twitter's emphasis on freedom of speech, Twitter does not remove this content, but simply reduces its visibility and notifications associated with it for tagged users.

[142] Gorwa, R., Binns, R. and Katzenbach, C. (2020) "*Algorithmic content moderation: Technical and political challenges in the automation of platform governance,"* Big data & society, 7(1), p. 205395171989794.

Table 4:3 Moderation system by issue area

	Terrorism	Violence	Toxic Speech	Copyright	Child Abuse	Sexual Content	Spam and automated accounts
Facebook	Shared Industry Hash Database (SIHD), ISIS/Al-Qaeda classifier	Community standards classifiers	Community standards classifiers	Rights manager	PhotoDNA	Non-consensual intimate image classifier, nudity detection	Immune system
Instagram	-	-	Comment filter	Rights manager	PhotoDNA	-	Content filter, false account detection CG ML Classifiers
YouTube	SIHD, Community Guidelines (CG) ML classifiers	CG ML Classifiers	CG ML Classifiers	Content ID	Content safety API, PhotoDNA	CG ML Classifiers	CG ML Classifiers
Twitter	SIHD	-	Quality filter	-	PhotoDNA	Sexual content interstitial	Proactive Tweet and account detection, quality filter
WhatsApp	-	-	-	-	PhotoDNA	-	Modified immune system

Source: Gorwa et al. (2020)

The core issues across the major platforms and the role of human moderation were also presented in Gorwa et al.'s paper:

Table 4:4 Notable algorithmic moderation systems

Actor	System	Issue areas	Target content	Core tech	Human role
YouTube	Content ID	Copyright	Audio, video	Hash-matching	Trusted partners upload copyrighted content
Google Jigsaw	Perspective API	Hate speech	Text	Prediction (NLP)	Label training data and set parameters for predictive model
Twitter	Quality filter	Spam, harassment	Text, accounts	Prediction (NLP)	Label training data and set parameters for predictive model
Facebook	Toxic speech classifiers	Hate speech, bullying	Text	Prediction (NLP, deep learning)	Label training data and set parameters for predictive model; make takedown decisions based on a flag
GIFTC	Shared-industry hash database	Terrorism	Images, video	Hash-matching	Trusted partners suggest content, firms find/ add content to a database
Microsoft	PhotoDNA	Child safety	Images, video	Hash-matching	Civil society groups/law enforcement add content to a database

Source: Gorwa et al. (2020)

4.2.3 Role of community and moderation in enforcing behaviour

Moderation exists at two levels[143]:

- **Community-driven (or self-):** Community-driven or self-moderation is managed internally by community appointed moderators who review user-generated content within an individual website community to ensure member content meets set rules and guidelines specific to this group, e.g., Wikipedia, Facebook (private group setting), and Reddit; and
- **Corporate (platform):** Platform moderation monitors all content generated on an individual site, and as outlined previously can be supported by human moderation teams or AI-driven solutions that flag potentially harmful content for review. e.g., Twitter and Instagram.

Seering (2020)[144] proposes that community or self-moderation occurs on three levels:

- *on an everyday, in-the-moment level,* moderators interact with community members, warn potential offenders and explain rules, remove content or users when necessary, and deal with the fallout of these removals;
- *on a level that spans weeks or months*, moderators learn how to moderate (this includes their processes of recruitment, role differentiation, learning how to handle various situations, and development of an overall moderation philosophy); and
- On *the broadest level*, which spans the entire lifetime of a community, moderators respond to internal community dynamics, platform developments, and cultural shifts by revising community rules and how they are enforced.

Seering also notes that while there are known issues with self-moderation (i.e., manual moderation of a group can support the emergence of a positive or toxic environment, dependent on community standards and individual moderator) this type of moderation should be understood and used in conjunction with wider platform moderation.

Well managed community moderation is known to have a range of positive effects, such as increased contribution and quality of input to the group, improved quality of user, and increased stability during difficult times.

Similar effects have been seen in AI moderation if the platform's algorithm is biased towards displaying a post's comments by perceived quality, which therefore promotes more quality engagement.

---

143 Seering, J. (2020) "*Reconsidering self-moderation: The role of research in supporting community-based models for online content moderation*," Proceedings of the ACM on human-computer interaction, 4(CSCW2), pp. 1–28.
144 ibid

Poorly managed communities can have wider negative impacts. This has been seen in the past with Facebook, where unregulated private groups are set up and managed by conspiracy theorists and extremists, which in turn created a pathway for radicalisation on the platform.[145]

This issue is prevalent on the network, which has tens of millions of groups, and Facebook has recently rolled out a new tool to support admins in assessing the quality of groups and in limiting the number of comments from potentially toxic users.[146]

As the number of private groups increases, and with the increased awareness of the role of private communities in radicalisation and the dissemination of misinformation, it is clear that there is the need for a solution that can moderate both at the platform and the community level.

Intervention at this level will be essential; Facebook data reveals that private groups relating to conspiracy theory QAnon hosted 3m members on the network. Of these users, 2m were attracted to groups based on recommendations made by the Facebook algorithm, which shows the extent of algorithmic amplification on the network.[147]

---

145 Paul, K. (2019). *Facebook's crackdown on dangerous content in groups could backfire, experts say*. [online] the Guardian. Available at: https://www.theguardian.com/technology/2019/aug/14/facebook-private-groups-rules-extremist-fake-news [Accessed 18 Oct. 2021].

146 Perez, S. (2021). *Facebook rolls out new tools for Group admins, including automated moderation aids*. [online] TechCrunch. Available at: https://techcrunch.com/2021/06/16/facebook-rolls-out-new-tools-for-group-admins-including-automated-moderation-aids/ [Accessed 18 Oct. 2021].

147 Breland, A. (2020). *Facebook announces crackdown on QAnon, antifa, and militias*. [online] Mother Jones. Available at: https://www.motherjones.com/politics/2020/08/facebook-qanon-antifa-militia/ [Accessed 18 Oct. 2021].

4.3 Validating what technology works

4.3.1 Introduction

This section sets out identified best practice that can inform future partnerships, setting out what an intervention should set out to do, avenues for engagement, and how to create a transdisciplinary product that is transparent and ethical.

4.3.2 Role of the Intervention

A report conducted by the Centre for Research and Evidence on Security Threats (CREST)[148] provides an overview of best practices within existing literature related to intervention and educating, highlighting the currently limited evidence base.

The report highlights that the key factors that drive disengagement from radicalisation pathways are disillusionment and wider social-ecological factors:

- **Disillusionment with a cause** can be driven by a range of factors, including frustration with group leadership, lack of progress in achieving stated goals, burnout, the group's inability to meet "core needs" (e.g., such as the search for identity) that motivated initial engagement, and concerns around the use of violence against civilians or other group members.
- **Socio-ecological and contextual** factors are also important, with previously radicalised individuals stating that the absence of a relationship with family members and friends outside of the movement makes it less likely for an individual to disengage. Interventions that address socio-ecological factors are increasingly used today and are showing positive results.

Counter-messaging campaigns have also been used, with varying success, and CREST notes that a systematic review suggests that they are more practical at earlier stages of radicalisation.

Best practice identified by CREST includes offering tailored multi-agency interventions designed to address the specific individual and ecological risk factors identified in individual cases. There is also promise in gendered intervention.

148 Lewis, J. and Marsden, S. (2021). *Countering Violent Extremism Interventions: Contemporary Research*. [online] Centre for Research and Evidence on Security Threats. Available at: https://crestresearch.ac.uk/resources/countering-violent-extremism-interventions/ [Accessed 18 Oct. 2021].

Key recommendations offered by CREST when designing intervention include:

1. Engagement with reformed radicals;
2. Working across agencies to develop social-ecological interventions;
3. Encouragement of community engagement and reporting of vulnerable persons;
4. Support the investigation into the effectiveness of online intervention;
5. Use of holistic tools to assess and manage risks;
6. Develop an in-depth understanding of the nuances that exist between and within different ideologies.

4.3.3 Different types of engagement

Commercial Many larger platforms provide support to start-ups in the form of venture capital, private equity funding and more expansive guidance programmes. For example, Facebook's Accelerator for Start-ups programme has so far supported 35 separate investments, two diversity investments and 89 acquisitions.[149] Example start-ups funded by Facebook that have potential applications within the SafetyTech sector include Lesan, an instant translation tool for low-resource languages. This can be used to promote access to the web for minority groups and may also have application in identifying harmful speech across languages.[150]

Funding is also available through organisations like End Violence Against Children who have financed the Internet Watch Foundation's reTHINK chatbot that engages with internet users who may be looking for images of child sexual abuse; providing early intervention and signposting service.[151]

Academic Increased understanding of how an individual acts online highlights the need for a multidisciplinary and evidence-based approach. In the UK research is ongoing with researchers from various backgrounds engaging in an £8.6m UKRI Research Centre of Excellence project focussed on the protection of citizens online.

[149] Mehrey, A. and Bharath (2021*). Facebook Startup Funding | Startups Funded by the Facebook*. [online] StartupTalky. Available at: https://startuptalky.com/facebook-funded-startups/ [Accessed 18 Oct. 2021].
[150]Gibbons, V.-M. (2020). *Celebrating Startup Success at Facebook Accelerator London*. [online] Facebook for Developers. Available at: https://developers.facebook.com/blog/post/2020/03/20/celebrating-startup-success%20-facebook-accelerator-london/ [Accessed 18 Oct. 2021].
[151]IWF. (2020). "*Game-changing" chatbot to target people trying to access child sexual abuse online*. [online] Available at: https://www.iwf.org.uk/news/game-changing%E2%80%99-chatbot-to-target-people-trying-to-access-child-sexual-abuse-online [Accessed 18 Oct. 2021].

The REPHRAIN project[152] aims to address fundamental tensions and imbalances pertaining to protecting citizens online through three overarching missions (to deliver privacy at scale while mitigating its misuse to inflict harm, to minimise harms while maximising benefits from a sharing-driven digital economy, and to balance individual agency vs. social good). REPHRAIN's missions are derived from two sources, the UK government's Online Harms White Paper and the University of Pennsylvania's Taxonomy of Privacy,[153] which includes a strand explicitly focused on promoting data sharing and availability.

Hatelab is a further example of a global, academic-led project focused on gathering data and gaining insight into hate speech and crime. Like REPHRAIN, Hatelab is funded in part by the UKRI, in partnership with the Economic and Social Research Council and the US Department of Justice. Through the project, the Online Hate Speech Dashboard has been developed between academics with policy partners. This has allowed the project to provide aggregate trends in hate speech and activity over time and space.[154]

Public / Networks The Safety Tech Innovation Network has been set up in the UK to support the promotion, collaboration, and industrial application of online safety technologies in the UK and further afield. The network encourages innovation by enabling more efficient collaboration on technical solutions by supporting information sharing, problem-solving and solution creation. It also seeks to drive the adoption of technology by showcasing the advantages and opportunities that come with safety technology, and by unifying and creating a consistent voice for the sector.

The Safety Tech Innovation Network has wider links to other organisations such as the GIfCT, WeProtect and the Fair Play Alliance and is funded by DCMS, Nominet and Innovate UK, delivered by KTN.[155]

Other examples of industry networks working to address online harms include the Online Safety Tech Industry Association (OSTIA) who are working together to ensure policymakers and larger companies are aware of innovation, technology,[156] and best practice for online safety, and the Trust and Safety Professional Association who work globally to develop a shared community of practice and definitions around online harm.[157]

---

[152] Rephrain. (2020). *National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online*. [online] Available at: https://www.rephrain.ac.uk/.
[153] Solove, D.J. (2006). *A Taxonomy of Privacy.* [online] University of Pennsylvania Law Review. Available at: https://scholarship.law.upenn.edu/cgi/viewcontent.cgi?article=1376&context=penn_law_review.
[154] Hatelab. (n.d.*). HateLab – A global repository for data and insight into hate crime and speech*. [online] Available at: https://hatelab.net/ [Accessed 18 Oct. 2021].
[155] Curtis, A. (2020). *About the Safety Tech Innovation Network. SafetyTech Innovation Network*. [online] 15 Oct. Available at: https://www.safetytechnetwork.org.uk/articles/about-the-safety-tech-innovation-network/ [Accessed 18 Oct. 2021].
[156] OSTIA. (2021). *OSTIA - Online Safety Tech Industry Association*. [online] Available at: https://ostia.org.uk/.
[157] Trust & Safety Professional Association. (2021). *Advancing the trust and safety profession through a shared community of practice*. [online] Available at: https://www.tspa.info/ [Accessed 18 Oct. 2021].

Open source MacAvaney et al.'s[158] (2019) paper reviews existing datasets that can be used to flag hate speech. They note the limitations across datasets, including the variation in hate speech definitions and the lack of transparency around decisions made by trained models.

They also note that typically there are not many publicly available (granular) datasets. Those identified in the study are included in the appendix, supplemented with additional sources identified in Yin and Zubiaga's (2021) study[159]. It should be noted that there are wider issues with open-source material, e.g., material may not be maintained or up to date, and there may be potential known and unknown biases within data. As hate speech is constantly evolving and changing the extent to which open-source models work effectively long-term is unknown.

A full overview of identified open data sources is outlined in the appendix.

4.3.4 Supporting transparency, efficacy and effectiveness in model design

A systematic review conducted by Yin and Zubiaga (2021)[160] reviewed all identified papers that use natural language processing to identify hate speech. The factors identified that impact the efficacy of a dataset includes:

- Search terms used when identifying hate speech;
- Topics covered;
- Labelling definitions;
- Data source platforms;
- Level of bias within the dataset;
- The proportion of abusive posts in the dataset; and
- Size of the dataset.

Current issues with NLP identifiers outlined within the paper and practical solutions offered by developers are also outlined below.

Table 4:5 Limitations and solutions within Natural Language Process Solutions

Problem	Solution
Most models struggle to identify similar levels of hate speech when dealing with new data.	• Further masked language modelling pre-training on an abusive corpus, or by incorporating features from a hate speech lexicon may support the development of generalised models.

158 MacAvaney, S. et al. (2019) "*Hate speech detection: Challenges and solutions*," PloS one, 14(8), p. e0221152.
159 Yin, W. and Zubiaga, A. (2021) "*Towards generalisable hate speech detection: a review on obstacles and solutions,"* PeerJ. Computer science, 7(e598), p. e598.
160 Yin, W. and Zubiaga, A. (2021) "*Towards generalisable hate speech detection: a review on obstacles and solutions*," PeerJ. Computer science, 7(e598), p. e598.

	• Multitask training can support generalisation across datasets. Multiple labelling approaches can identify where labels overlap. • Less specific labelling may also help generalise models. (e.g., toxicity, offensive, abusive).
Most models are trained through English (5 models identified that operate in other languages)	• *Support the development of multilingual datasets/ development of global multi-lingual resource* • *Support the development of translation models that incorporate terms from the hate speech lexicon.*
Natural language processors perform best when trained and applied to specific domains	• *Facilitate the development of flexible resources that can be incorporated into different tools that are platform-specific tools.*
User grammar styles create an issue for NLP and can be used to evade detection. Training systems to identify euphemisms and additional spellings also have the potential to flag false positives.	• There are a range of methods that build relationships between words and accounts e.g., identifying hate communities through connections with extremist articles and their authors, e.g., FastText measures similarity and relatedness in words and the word2vec model that captures the similarity between words. • Other solutions to non-standardised grammar and vocabulary include the analysis of character-level features at the character and word level; sentence embedding which looks at the whole sentence in context and not the individual words; and the use of larger language models which look at sub-word and sentence embedding.
Small datasets can lead to overfitting and harm generalisability. Labelling is also challenging as hate speech is generally subjective.	• The use of pre-trained embeddings has been accepted as standard practice in the field of NLP to prevent overfitting and are common in hate speech detection as well. • **However,** the effectiveness of domain-general embedding models is questionable, and there has been only a limited number of studies that investigate the relative suitability of different pre-trained embeddings on hate speech detection tasks.

Non-random sampling and subjective annotations introduce individual biases, and the different sampling and annotation processes across datasets further increase the difficulty of training models that can generalise across heterogeneous data.	● Different annotation and labelling criteria result in essentially different tasks and different training objectives, despite their data having a lot in common. ● As a result of the varying and sampling methods, definitions, and annotation schemes, what current models can learn on one dataset is specific to the examples in that dataset and the task defined by the dataset, limiting the models' ability to generalise to new data. ● Example biases include author bias and topic bias.
There may be a representation bias as models are trained on societal norms.	● Minority groups are underrepresented in available data. This can lead to false positives within minority groups. The prevalence of such biases means that existing hate speech detection models are likely to struggle at generalising to unseen data that contain expressions related to these demographic groups. ● Compared to the other biases mentioned above, they do more harm to the practical value of the automatic hate speech detection models. These biases may cause automatic models to amplify the harm against minority groups instead of mitigating such harm as intended. ● Potential solutions to this representation bias include the incorporation of minority classifiers into modelling to de-bias the model and to re-train false positives.

Source: Yin and Zubiaga (2021)

In addition to the above Escartín et al. (2021)[161] conducted a survey of researchers, relating specifically to NLP shared tasks, producing a Shared Task Organisation Checklist. These considerations cover four key themes: transparency, reporting and replicability, system ranking, and metrics and are outlined in full in the appendix of this report.

[161] Trotzek, M., Koitka, S. and Friedrich, C. M. (2020) "*Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences*," IEEE transactions on knowledge and data engineering, 32(3), pp. 588–601.

4.3.5 Supporting the development of transdisciplinary teams

Cho and Kwon's (2015)[162] study into the effect of voluntary and policy-driven ID verification provides a good example of why the incorporation of behavioural scientists into a product team is important.

Their study found that harmful comments were significantly reduced when ID verification was introduced to social media on a voluntary basis, especially among moderate service users. For mandatory, policy-driven verification however hate speech did not change and, in some situations, increased.

Sahneh et al. (2021)[163] provide a list of rules that help cultivate transdisciplinary teams within data science, developed based on the authors' experience working in previous multidisciplinary teams in the past. This list is included in full in the appendix of the report, and potential ways the foundation can support this process are outlined below:

- Support the development of a transdisciplinary toolkit to include core concepts and terms (e.g., overview of regulation, technical language, and a lexicon of hate) from relevant disciplines (psychological, hate speech-specific, political, technological);
- Support best practice during the planning process (defining the purpose of the collaboration, assigning roles and responsibilities etc.);
- Ensure the project follows FAIR data principles (Findable, Accessible, Interoperable, and Reusable) and that written code is understandable and transferable;
- Support the integration of ethical and wellbeing perspectives at each stage of the project; and
- Support peer-learning by establishing a network that allows practitioners to share best practice and resources to aid future transdisciplinary working.

[162]Cho, D. and Kwon, H. (2015) *The impacts of identity verification and disclosure of social cues on flaming in online user comments*, Researchgate.net.

[163] Sahneh, F., Balk, M.A., Kisley, M., Chan, C., Fox, M., Nord, B., Lyons, E., Swetnam, T., Huppenkothen, D., Sutherland, W., Walls, R.L., Quinn, D.P., Tarin, T., LeBauer, D., Ribes, D., Birnie, D.P., Lushbough, C., Carr, E., Nearing, G. and Fischer, J. (2021). *Ten simple rules to cultivate transdisciplinary collaboration in data science*. PLOS Computational Biology, 17(5), p.e1008879.

4.4 Emerging trends in Safety Tech

4.4.1 Market disruption and changing online trends

This section looks at the changing online landscape relevant to Safety Tech, assessing what this means for the sector. Specifically, it outlines current engagement with AI, the emergence of video-centric and streaming platforms, wider changes in social media habits, and the continued rise of hate speech and misinformation.

Artificial Intelligence The use of AI and machine learning to support human moderation is one of the key developments within Safety Tech in recent years. Ofcom's (2019) report[164] assesses the impact of AI in online content moderation, outlining its role at the pre-moderation and moderation stage.

At the pre-moderation stage, hashing and keyword filtering techniques can flag potentially negative content, but these techniques face challenges associated with grammar, evolving language, emojis and pictures. Emerging techniques that have the potential for addressing these issues include object detection, scene understanding, natural language processing and sentiment analysis. The incorporation of metadata can also play a role. Metadata can include a wide variety of information such as the user's post history, friends, age, location etc.

AI has the potential to support existing human moderators and can prioritise images based on the level of or ambiguity of harm. It can also blur out identified harmful content to limit exposure for moderators.

As outlined previously computer vision is set to mature within the next 2-5 years. AI training is ongoing across major social networks such as Facebook who are working internally and with subcontractors, conducting sometimes exploratory work at the forefront of the technology development[165, 166].

This raises concern around the transparency and efficacy of internal projects and the role of subcontractors within this.

---

[164] OFCOM (2019) *Use of AI in Online Content Moderation www.ofcom.org.uk*. Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf (Accessed: September 17, 2021).
[165] Bernal, N. (2021) *Facebook's content moderators are fighting back*, *WIRED UK*. Available at: https://www.wired.co.uk/article/facebook-content-moderators-ireland (Accessed: September 17, 2021).
[166] Buntz, B. (2020) *2020 predictions: Computer vision projects will gain ground*, *Iotworldtoday.com*. Available at: https://www.iotworldtoday.com/2020/01/06/2020-predictions-computer-vision-projects-will-gain-ground/ (Accessed: September 17, 2021).

Use of subcontractors to train AI Facebook outsources content moderation services globally. Sub-contracted staff in Ireland have raised concerns about their treatment by the firm, stating they are unfairly paid, that they coerced into signing non-disclosure agreements, that if they underperform, they receive a cut in pay, and that the support they receive after their exposure to harmful and illegal content is minimal.[167,168]

Facebook now faces pressure in Ireland to update its outsourcing model, which in turn may have implications for other firms operating similar models. Campaigners are currently calling on Facebook to directly employ moderators and address the *"cynical attempt to deny them rights in this day and age"*.

Sub-contractors are also accused of using non-disclosure agreements to limit the extent to which sub-contracted employees can discuss the standard of work conditions.

It should be noted that Facebook competitor TikTok has recently hired over 500 in-house moderators in Ireland, which in turn may pressure Facebook into updating its approach to content moderation. TikTok is also in the process of shifting its moderation to more local teams so nuance across regions can be better understood. It is also working with US law firm KandL Gates to increase its transparency in the United States.[169]

Use of subcontractors to develop solutions As noted, many larger platforms work with subcontractors to develop solutions that tackle hate speech and other online harms. While this can have a positive impact on the main social media platforms' capability to address harms, it is important to consider the extent to which these partnerships are effective and ethical.

Veteran fact-checking platform Snopes outlined that they would not work with Facebook due to how the social network addressed false and misleading media.

The firm's vice-president stated that ongoing work internally at the larger platforms was "credibility theatre" and that *"The fact that Facebook has more people on their PR staff than there are formal fact-checkers in the world demonstrates the disproportionality of the situation."*[170]

---

[167] Bernal, N. (2021) *Facebook's content moderators are fighting back*, *WIRED UK*. Available at: https://www.wired.co.uk/article/facebook-content-moderators-ireland (Accessed: September 17, 2021).
[168] Newton, C. (2019) *Facebook moderators break their NDAs to expose desperate working conditions*, *The Verge*. Available at: https://www.theverge.com/2019/6/19/18681845/facebook-moderator-interviews-video-trauma-ptsd-cognizant-tampa (Accessed: September 17, 2021).
[169] Murphy, H. and Yang, Y. (2019) "*TikTok rushes to build moderation teams as concerns rise over content," Irish times, 20 December*. Available at: https://www.irishtimes.com/business/technology/tiktok-rushes-to-build-moderation-teams-as-concerns-rise-over-content-1.4121460 (Accessed: September 17, 2021).
[170] Coldewey, D. (2019*) "Snopes rolls its own crowdfunding infrastructure to prepare for 2020's disinformation warfare*," *TechCrunch*, 20 December. Available at: http://techcrunch.com/2019/12/20/snopes-rolls-its-own-crowdfunding-infrastructure-to-prepare-for-2020s-disinformation-warfare/ (Accessed: September 17, 2021).

An independent report conducted by Snopes went on to identify a network of fake profiles coordinated to support right-wing president Donald Trump and operating out of Vietnam, reporting that Facebook did not address the network prior to the publication of the report and have since taken no further action.[171]

It is therefore important to consider how major social platforms monitor and act upon known harm and consider how activity on larger firms can be evaluated and held to account externally.

Social media habits The Pew Research Centre conducted a survey that focused on social media habits.[172] Usage statistics are presented below:

Table 4:6 Social Media Usage

Social Network	Percentage of US population using the platform
YouTube	81%
Facebook	69%
Instagram	40%
Pinterest	31%
LinkedIn	28%
Snapchat	25%
Twitter	23%
WhatsApp	23%
TikTok	21%
Reddit	18%
Nextdoor	13%

Source: Pew Research Centre

[171] Snopes (2019) *If Facebook is dealing with deceptive 'BL' network, it's not working Snopes.com*. Available at: https://www.snopes.com/news/2019/12/13/facebook-bl-cib/ (Accessed: September 17, 2021).
[172] Atske, S. (2021a) *Social media use in 2021*, *Pewresearch.org*. Available at: https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/ (Accessed: September 17, 2021).

Survey findings show that YouTube and Facebook are the main social networks and are used by the majority of the population surveyed. When comparing data from the previous year the two sites that showed significant growth were Reddit and YouTube (note TikTok and Nextdoor were only included in the most recent iteration). The use of other networks has remained consistent across surveyed years.

The platforms most popular with young people (18–29-year-olds) include Instagram, Snapchat and TikTok, each of which is largely image or picture-based.

Streaming platforms and the video medium The growth of streaming platforms in recent years must be considered going forward. Whereas platforms such as Twitter, Facebook and to an extent Instagram are driven by written communication, growth in platforms such as Twitch and TikTok has presented new issues for online safety firms.

In 2019 it was reported that TikTok was downloaded c.1.5bn times by November 2019[173] and Twitch has c.1.38bn concurrent users.

The development and popularisation of the above networks will mean that Safety Tech will likely have to expand to encompass not just NLP for words but to develop computer vision for mass media moderation.

Increased risk of hate speech Rise in hate speech was reported by the Report Harmful Content group[174], which reported a rise from 19 reports in 2019 to 64 reports in 2020. This increase in hate speech is reflected in Digital Awareness UK's report which states that despite positive support to the Black Lives Matter movement there has also been a *"marked increase in posts encouraging harmful ideologies, such as anti-immigration, homophobia, xenophobia, racism and anti-Semitism."*

Increased risk of misinformation An analysis conducted by NewsGuard revealed that interaction with unreliable news sources through social media doubled between 2019 (8% of all interactions) and 2020 (17% of all interactions), while general engagement with media in general through social media increased from 8.6bn to 16.3bn.[175] This suggests that the use of social media as a news source is becoming increasingly normalised and that there is a need to properly moderate and act upon instances of misleading or false news articles.

[173] Onix (2021) *How TikTok has changed live streaming for social media*. *Onix-systems.com*. Available at: https://onix-systems.com/blog/how-did-tiktok-social-media-live-streaming-change-everything (Accessed: September 17, 2021).
[174] Kathryn Tremlett (2021) *A sit down with report harmful content*, *Org.uk*. Available at: https://swgfl.org.uk/magazine/a-sit-down-with-report-harmful-content/ (Accessed: September 17, 2021).
[175] Stewart, E. (2020) *America's growing fake news problem, in one chart*, *Vox*. Available at: https://www.vox.com/policy-and-politics/2020/12/22/22195488/fake-news-social-media-2020 (Accessed: September 17, 2021).

There is also a need to address the current algorithm design, which as mentioned above is biased towards reactive platform use, promoting "engaging" posts, which are often controversial.

Increased risk of radicalisation Two key trends identified in the research include the emergence of lone actors and the ease of online radicalisation.[176] There are also trends that a higher proportion of women are radicalised than before on the internet.[177]

The current approach to combat this within the EU is outlined in the Terrorist Content Online Regulation (2021)[178] which require social media companies to remove harmful content within one hour of the notification period. As part of this regulation, social media firms are called on to provide annual transparency reports on measures taken to remove terrorist content.

Movement of alt-right to alternative sites It is important to note the movement of alt-right social media users to alternative platforms. While platforms such as MeWe may not set out to attract the alt-right (MeWe describes itself as a "privacy-first" social network that does not promote content, sell advertising space, or harness user data.) their approach to moderation can make them an ideal network for alt-right media. MeWe saw growth in users from 12m at the end of 2020 to 16m in February 2021, which was driven by the disillusionment associated with the US election.[179]

Similar significant growth was noted on Gab – the alt-right Twitter alternative which saw 1.7m users signing up one week after. Rumble, a YouTube alternative doubled in the same week, and the Telegram app reached number 2 on the Apple download charts. While using these networks may not be problematic the lack of regulation can have real-world negative results. As an example, Telegram has been used as an effective recruitment tool by Isis during its wars in Iraq and Syria.[180]

---

[176] Pandith, F. and Ware, J. (2021) *Teen terrorism inspired by social media is on the rise. Here's what we need to do*, *NBC News*. Available at: https://www.nbcnews.com/think/opinion/teen-terrorism-inspired-social-media-rise-here-s-what-we-ncna1261307 (Accessed: September 17, 2021).
[177] Pandith, F., Ware, J. and Bloom, M. (2020) *Female extremists in QAnon and ISIS are on the rise. We need a new strategy to combat them*, *NBC News*. Available at: https://www.nbcnews.com/think/opinion/female-extremists-qanon-isis-are-rise-we-need-new-strategy-ncna1250619 (Accessed: September 17, 2021).
[178] European Commission (2021) *Terrorist content online*, *Europa.eu*. Available at: https://ec.europa.eu/home-affairs/system/files/2021-05/202104_terrorist-content-online_en.pdf (Accessed: September 17, 2021).
[179] Bond, S. (2021) "*Fast-Growing Alternative to Facebook and Twitter Finds Post-Trump Surge 'Messy,'*" NPR. Available at: https://www.npr.org/2021/01/22/958877682/fast-growing-alternative-to-facebook-twitter-finds-right-wing-surge-messy?t=1630676748964 (Accessed: September 17, 2021).
180 Ray, S. (2021) "*The far-right is flocking to these alternate social media apps — not all of them are thrilled*," Forbes Magazine, 14 January. Available at: https://www.forbes.com/sites/siladityaray/2021/01/14/the-far-right-is-flocking-to-these-alternate-social-media-apps---not-all-of-them-are-thrilled/ (Accessed: September 17, 2021).

5 Theme 3: Economics

5.1 Introduction

This section looks at the economics of online hate, exploring the incentives to create, host, share or prevent hate across varying parties. It is different from the previous sections, in that it explores the incentives that are in place and underpinning current online hate and harms, as well as those incentives to address, mitigate or remove hateful content and behaviour.

Key findings from this section include:

- Incentivisation is important for the creation and proliferation of hateful content online. Poor platform design can incentivise the creation of more controversial content, and revenue generated from this may further encourage problematic behaviour or directly fund hate groups.
- Platforms and advertisers claim that hateful content is not in their interest, but their engagement-based revenue models are known to encourage hate speech;
- There are substantial market and investment opportunities in areas such as social media management and analytics (expected CAGR of 24%), threat intelligence (CAGR of 19%), and counter-disinformation (CAGR of 42%). Growth is strong but less pronounced in areas such as content moderation (CAGR of 10%) and filtering and parental control (CAGR of 11%).
- Potential areas of growth, or increased interest include:
    - The identification of harmful actors
    - The identification of hate speech and disinformation
    - Embedding trust and safety into the wider technology ecosystem
    - Solutions that introduce real-world consequences into online hate
    - Intervention-based approaches
    - Solutions that monitor hate across platforms and
    - Innovation within the content moderation subsector.
- The average time taken to secure an initial investment among safety tech businesses has fallen significantly in recent years suggesting an increasing demand and need for seed investment within the sector. Further, the encouragement of academic start-ups within this domain may also enable investors to back particularly early-stage / stealth start-ups.

5.2 The role of incentives

This section of the report outlines how incentives work across different elements of the online harm landscape, and where the opportunities exist at the individual, platform, and regulatory level.

The individual A blogger discussing disinformation online/ creating hateful content provides a good example of how economic incentivisation can influence behaviour. While the individual may not be fully aware, to begin with, they may increasingly share or engage with hateful content if their revenue or subscription rate increases alongside it. If this revenue stream is consistent and their behaviour is reinforced, this economic incentivisation can result in a sustained pipeline of disinformation within the broader information environment, ultimately normalising hate speech and marginalising targeted groups.

The platform The platform that hosts this content may have an economic incentive to continue to host that content, as it drives engagement with their platform, increasing advertising revenue streams and increasing their user base. In this example, there is a mutual dependency and relationship between the individual and the provider to create, host, and share problematic material.

The platform must therefore consider a potential trade-off: continue to host the content with associated potential financial benefits or remove the content with potentially negative economic effects such as loss of users or revenue.

Despite this potential for economic trade-offs, many platforms have insisted that they do not benefit from hosting hateful material[181] - even though the nature of their business model requires engagement that can be facilitated by harmful content i.e.

"I want to be unambiguous: Facebook does not profit from hate. Billions of people use Facebook and Instagram because they have good experiences — they don't want to see hateful content, our advertisers don't want to see it, and we don't want to see it. There is no incentive for us to do anything but remove it." Nick Clegg, VP Meta

"*There's also a broader problem: Facebook's advertising-based business model is powered by engagement—its algorithm promotes whatever content keeps people hooked. The system makes Facebook the perfect breeding ground for conspiracies and disinformation of all sorts, including climate denialism, because that kind of content is some of the most engaging…. Facebook makes money by luring users to the platform and keeping them on the platform. If people are spending less time on*

181 Clegg, N. (2020). *Facebook Does Not Benefit from Hate*. [online] About Facebook. Available at: https://about.fb.com/news/2020/07/facebook-does-not-benefit-from-hate/.

the platform, [Facebook] makes less money."[182] Danny Rogers, Global Disinformation Index and an adjunct professor at New York University

The role of economics in the online harms debate is an important one, in that it can help stakeholders in the area understand the incentives, the impact, and the response to online harm. Economic analysis can be used to consider the costs of action and inaction in tackling online harms and to evaluate the appropriateness and potential response from respective economic agents in regulatory scenarios.

Regulatory bodies This underpins the rationale for various governments introducing regulatory considerations for online harms, in that 'self-regulation' may often be outweighed by 'self-interest' – and that a government agent should intervene to set parameters for illegal and harmful content.

Further, whilst there is little dispute regarding the need to takedown evidently illegal material across jurisdictions, there are often costs and benefits associated with tackling 'harmful' material at a platform level. For example, if content is potentially harmful to one agent, the platform may need to consider their established terms and conditions, as well as consider the costs of identifying such material and responding (e.g., investment in moderation and response processes).

In equal measure, the introduction of legislation by governments to address online harms may also provide incentives for platforms and online providers to 'over mitigate' harmful content (e.g., over-moderation) which may, in turn, generate unintended impacts such as 'chilling effects' or encourage users to use alternative less-moderated platforms.

[182] de la Garza, A. (2021). *What Would a Climate-Conscious Facebook Look Like?* [online] Time. Available at: https://time.com/6100770/facebook-climate/ [Accessed 18 Oct. 2021].

5.3 The commercialisation of content creation

The increased scale and significance of online platforms has generated a wide range of market opportunities for individuals, businesses, and advertisers.

For example, a video-sharing platform requires creators to upload and generate content. To view this content, they may require the end-user to view an advert or pay for 'ad-free' access to generate revenue, and to generate engagement, they will pay the content creator a proportion of the revenues to generate further content and so on.

In other words, the platform has an incentive to increase engagement to increase advertising revenues. Advertisers have an incentive to engage with high-engagement channels. The content creator has an incentive to generate content that will generate high engagement and associated revenue. The individual also has an incentive to engage with content that is easy to access, 'low-cost', and of interest to them.

This interdependency between advertisers, platforms, content creators, and the individual creates a multitude of incentives at all levels, which left unchecked may cause problematic behaviours.

The commercialisation of content creation means that hateful or harmful content (left unchecked) can be monetised, even inadvertently and unknowingly to those providing the underlying funding. This might include:

- *An individual shares disinformation about COVID-19 and vaccinations on established social media platforms*. In response to content moderation or temporary bans, they claim they are being silenced by the establishment. They set up an alternative blog-site, and request monthly donations through a payment site, thereby monetising disinformation, and incentivising further generation of false content. Negative impacts generated could include lower vaccination take-up and worsened health outcomes/death, funding of disinformation, increased civil unrest, or threats towards public health officials.
- *Extremist groups organising and receiving funding through online platforms:* Despite recent attempts to curtail extremist activity on mainstream social platforms, internal Facebook research suggested that in 2020, there were up to 3 million followers of QAnon content[183]. Whilst such activity can be responded to, this may result in users following such groups onto alternative platforms, and further – provide revenue streams for such groups through mailing lists and donations. The GDI has identified at least 54 funding mechanisms for such

[183] Sen, A. and Zadrozny, B. (2020). *QAnon groups have millions of members on Facebook, documents show*. [online] NBC News. Available at: https://www.nbcnews.com/tech/tech-news/qanon-groups-have-millions-members-facebook-documents-show-n1236317.

groups including retail, donations, cryptocurrencies, content subscription sites, crowdfunding, and direct requests.[184]

This suggests that disincentivising hateful content through mechanisms such as friction and, ultimately, defunding may act as powerful tools. However, this requires three main components by a wide array of platforms:

- **Identification of hateful content and behaviour** (this can be supported by the development and implementation of safety tech solutions)
- **Knowledge sharing** – the online hate landscape contains a range of bad actors, varying lexicons and platforms in usage, and targets. The use of knowledge sharing or intelligence regarding these groups and terms may enable platforms to detect extremist or hateful content at an early stage. This might take the form of technical solutions (e.g., identifying extremist groups using alternative platforms, and organising to disrupt this) or through broader research and collaborative engagement across the trust and safety ecosystem e.g., sharing lexicons, and what works. There are established models for knowledge sharing within cyber security (e.g. CiSP) which could be emulated for trust and safety.
- **Commitment (or requirement) to respond:** Perhaps most importantly, there needs to be a willingness to identify and respond to online hate where it can be found. The extent to which platforms may be willing to do so may be impacted by incentives to respond, and regulatory considerations. Further, the proportional cost (examined later within this review) may also be a core consideration.

Overall, this suggests that tackling online hate requires significant disincentives for the creation and sharing of such material and that platforms, advertisers, and governments need to consider their role in identifying the monetisation or growth of groups sharing such material.

[184] Disinformation Index. (2020). *Bankrolling Bigotry: An Overview of the Online Funding Strategies of American Hate Groups.* [online] Available at: https://disinformationindex.org/wp-content/uploads/2020/10/Bankrolling-Bigotry_GDI_ISD_October-2020L.pdf.

5.4 The market for Safety Tech

In recent years, research has been undertaken to explore the emergence of the 'safety tech' sector (e.g., the Safer Technology, Safer Users work mentioned previously). Whilst this is a nascent and emerging sector, there has been a substantive history (spanning almost thirty years) of technological approaches to keeping users safer online.

In the 1990s, this typically focused on web filtering, digital forensics, and parental controls. The emergence of social media platforms such as Myspace, Facebook and Bebo and video content sites in the early 2000s, also created a route for the development of content moderation and platform monitoring approaches.

In the last few years, there have been several factors that have further amplified the reach, offering, and size of the sector. These include (but are not limited to):

- An expanded scope driven by new forms of technology e.g., deep fakes, live streaming, combined with an increased user base and proliferation of content.
- The emergence of regulation internationally has required platforms and publishers to respond and remove forms of illegal and/or harmful content.
- The increasing volume of content online requires a move towards using data analytics and AI in the content moderation process
- An emerging market preference to minimise toxicity and abuse from platforms, and to disassociate platforms from harmful or damaging material
- Personal exposure to harm and privacy abuse has meant that some customers expect safety to be a core consideration within online products.

However, despite the nascence of the 'Safety Tech' sector, we can identify a number of distinct sub-markets (which can be 'pure-play' e.g., content moderation) or more diversified (e.g. identity verification, of which age assurance may be one component) that have shared ambitions under the 'safety tech' banner i.e., to protect users from harm.

The UK's Safety Tech sectoral analysis identified several distinct sub-sectors, including 'system-wide governance', 'platform governance', 'platform moderation and monitoring', 'age orientated online safety', 'user protection including user-initiated protection and network filtering' and 'information governance', as outlined in Theme 2 previously.

Example Estimates

Whilst the most recent UK study identifies approximately 100 dedicated firms, with revenues in excess of £300m, there is less information available underpinning the broader 'safety tech' market globally. However, initial estimates suggest the UK has approximately 25% global market share with respect to firm count.

Internationally, recent investment raised by firms such as ActiveFence, L1ght, Truepic, and Blackbird.AI suggests significant investor interest in areas such as content moderation, threat actors, disinformation, and brand protection, and detecting illegal and harmful content. This also suggests a density of activity in the United States, United Kingdom, Israel, and Europe.

The Alfred Landecker Foundation is currently working with Dealroom to undertake an international Safety Tech mapping exercise, which will be a highly useful asset for further understanding international activity.

Further, there are a number of broad estimates for sectors and technologies aligned to the safety tech definition. We have identified the following broad market studies (however, these have not been verified or tested):

- Content Moderation (estimated to reach a market size of c. $11.8bn by 2027, with a Compound Annual Growth Rate (CAGR) of 10% between 2019-2027). This suggests:
    - Media and entertainment, and retail and e-commerce businesses each account for a third of this revenue.
    - More than 60% of the end-users prefer content moderation services over software (suggesting demand for outsourcing, or APIs as a minimum.
    - "According to various estimates, there are more than 100,000 people involved in core moderation businesses for brands at any point in time. However, the deluge of user-generated content in real-time creates a significant gap".
- Identity Verification (currently estimated at $7.6bn in 2020 and expected to reach a market size of $15.8bn by 2025, i.e., a CAGR of 15.6%. This also suggests:
    - Significant market growth is driven by increasing digital initiatives, an increase in fraudulent activities, and identity theft.
    - This market contains a number of non-dedicated safety tech providers e.g., GBG, which work in age assurance and age verification – but across broader market verticals e.g., banking verification. However, identification of 'age assurance' specific technologies and vendors is an interesting and distinct sub-market, whereby expertise in 'under 13/under 18' regulations, data management, and data consent is needed.
- Counter-disinformation (a baseline estimate is not provided publicly; however, this suggests a CAGR of 42% between 2020-26. This rate of growth exceeds

that identified within the UK Safety Tech Sectoral Analysis (2020/21) of 35%, suggesting a particularly high-growth market opportunity). This also identifies:

  - Companies in this space are working on technologies and approaches including AI analytics that can identify synthetic media, Digital authentication solutions, content and social media monitoring tools (OSINT), Fake profiles detectors and fact-checking tools
  - Current market mechanisms mean that end users are debating whether to work with an external vendor or to rely on internally developed capabilities as well as existing OSINT monitoring tools re-directed to a dedicated team (explored further in the subsequent sub-sections)

- Kidtech advertising – the global children's digital advertising market is estimated to reach $1.7bn by 2021 (equivalent to 37% of all children's advertising spend). This research also suggests that:
  - More than 170,000 children go online for the first time every day;
  - More than 40% of new internet users in 2018 were children;
  - Privacy regulations such as GDPR-K and COPPA could help to protect more than 800m children globally online;
  - "Apple, Disney, Netflix and Amazon spend up to $3bn each year on high-quality children's content yet this lies behind a subscription paywall, unreached by advertisers". Ensuring that advertising is kidtech-enabled and compliant could be beneficial to the adtech sector.
- Social Media Management – this market is currently estimated at $14.4bn but is expected to reach $41.6 billion in 2026 (CAGR of 23.6%). Further:
  - This market can be segmented into "social media listening, monitoring, and analytics; social media asset and content management; and social media risk and compliance management"
  - There is an interesting overlap between these sub-markets, and broader threat, harm, risk and 'bad actor' intelligence (explored later).
- Digital Forensics – this market is estimated to grow from $4.2bn in 2017 to $9.7bn by 2022 (CAGR of 15.9%)
  - This is an important market for data identification, recovery, and analysis – particularly for illegal and harmful content. The proliferation of end-user devices and cloud means that the market is expected to grow substantially.
  - There are a number of important buyers within this market, including government, law enforcement, professional services, telecoms, and IT.
  - Further, there is some evidence that some providers in this market are starting to diversify their offerings to cover more tenets of online safety. For example, Oxygen Forensics' software includes facial and image recognition to identify harmful material such as weapons or nudity. In the UK, Cyan Forensics has launched 'Cyan Protect' to identify and moderate harmful content.

- Web Filtering – this market size was valued at $4.1bn in 2020 and is projected to reach $8.7bn by 2028 (CAGR of 11.5%). This includes DNS, keyword, and URL filtering, often used within a cyber security context, but also with application to education and children's devices (e.g., blocklisting harmful sites).
- Parental Control – this market is estimated at $900m in 2020 and is projected to reach £2.16bn by 2028 (CAGR of 11.6%). This research also suggests a somewhat fragmented market, with a number of acquisitions underway by providers to secure market share (e.g., SafeToNet's acquisition of NetNanny) across market verticals (e.g., education, mobile phone networks).
- Threat Intelligence – the 'Threat Intelligence' market was valued at $5.5bn in 2019 and is projected to reach $20.2bn by 2027 (CAGR of 19%). Whilst this market is typically pitched at understanding cyber security threats and vulnerabilities, there are a number of providers in this space engaging with datasets linked to online safety e.g., personal data leaks, identification of harmful groups or actors.
- AI Image Recognition Market – this market includes image and video recognition across various fields e.g., vehicle safety, advertising, security and surveillance, biometrics etc. This market was valued at $1.7bn in 2020 and is expected to reach $5.2bn by 2026 (CAGR of 25%). This market has, with respect to population surveillance in a safety and security context, been subject to some challenge. For example, the European Parliament has called for a ban on police use of facial recognition technology in public places, and on predictive policing.[185]

Overall, whilst these are very broad estimates of market activity, this suggests that:

- There are substantial market and investment opportunities across the safety tech domain. These appear most pronounced in areas such as social media management and analytics (expected CAGR of 24%), threat intelligence (CAGR of 19%), and counter-disinformation (CAGR of 42%).
- Growth is strong, but potentially less pronounced in areas such as content moderation (CAGR of 10%) and filtering and parental control (CAGR of 11%), suggesting these markets may be more defined, but also contain a number of businesses holding existing market share.

[185] HEIKKILÄ, M. (2021). *European Parliament calls for a ban on facial recognition*. [online] POLITICO. Available at: https://www.politico.eu/article/european-parliament-ban-facial-recognition-brussels/ [Accessed 18 Oct. 2021].

Market Models and Expenditure in Safety Tech

Whilst much of the revenue and market opportunity is set out within the previous estimates, it is important to consider two other factors within the safety tech sector.

Firstly, what are the market models and mechanisms for the development and implementation of safety within digital platforms? Secondly, what levels of expenditure are being sustained by different organisations?

Considering the first question, we have identified the following potential market approaches used by platforms internally or by independent safety tech providers.

Table 5:1 Safety Tech Market Models

Type	Examples	Definition	Benefits	Drawbacks
Free-to-use	PhotoDNA Kids Web Services Perspective API	Many providers offer 'free' or low-cost access to technical approaches. These often are offered by larger platforms to help address common problems (e.g., identifying and removing CSAM material)	Low/free cost to access should increase adoption. Entry-point into trust and safety techniques Externally validated / industry-standard approaches (e.g., PhotoDNA) used across platforms.	Not applicable to all domains of safety tech The potential risk of 'free' solutions used to reduce investment in more bespoke approaches by sector. Technology can be developed by larger organisations, therefore may require proprietary agreement / NDAs / adherence to T&Cs
Cloud provision (e.g., pay-per-use API)	Amazon Rekognition Azure Content Moderator	This includes the use of APIs or cloud solutions (pay per use) that can be integrated with existing tech stacks. For example, use of the Amazon Rekognition API to add pre-trained models for detecting nudity or toxic words.	Can be low-cost / scale with existing provision Pre-trained models (ease of use) to identify a large extent of harmful material e.g., extremist imagery	Underlying explainability of models / outcomes may be proprietary May require enhanced context (e.g., identification of 'bad words' without context'
Bespoke / in-house development	Twitter Safety Match	This refers to where existing platforms develop (or acquire) their own trust and safety solutions internally.	Models can be tailored to the platform's requirements Potential for development of in-	Can be resource-intensive Can have conflicting expectations within teams

	Group		house trust and safety teams Access to platform level data (understand context, creation, user T&Cs) Ability to acquire new technologies / start-ups	e.g., implementation of a trust feature may impact UI/UX Potential for commercial / restrictive deployment e.g., content moderation models or decisions not shared outside of the organisation (limits transparency / understanding of efficacy)
Outsourcing	Cognizant Accenture ModSquad	Whilst much content moderation can be done through automated processes, often human moderators are required to review the remainder of flagged content. Many of the largest social media platforms outsource their human-driven content moderation to third parties.	Can be done at scale (large scale recruitment). Addresses some gaps where human moderation / context is required.	Ethical / legal / health considerations – the human side of content moderation (set out previously). Roles often outsourced to low-income countries. Terms and conditions / processes for moderation not always well-defined (human subjectivity)
Third-party / independent provision	Covers the 'Safety Tech' sector (see Safety Tech	This refers to independent provision of safety tech solutions, whereby products or services are procured by a third party. This might mean, for example, an	Independent expertise can be developed and deployed across sector use cases, including private and public sector	Awareness remains an issue Efficacy / standards are not necessarily in place for some early-stage firms

	Providers)	advertising agency hiring a social media threat intelligence firm to identify bad actors, or potential threats to their brands online / a business hiring a digital forensics firm to scan devices for any illegal or harmful material in the event of investigation.	Clear resources available for end-customer Many market delivery mechanisms are available to embed trust and safety (e.g., filtering by default by ISPs for under-18s may reach more end-users than the use of parental controls at a device level).	Can be difficult to identify 'what works' given the nascence of the market. Firms need to establish a working business model (e.g., selling directly to the consumer may be less effective than selling B2B / to the public sector).
External hosting and processing	e.g., payment processors, data centres	Increasingly, the role of external actors involved in content hosting / payment processing has come under scrutiny. For example, many web hosting firms may have T&Cs regarding not posting harmful or hateful content, but may inadvertently host this content e.g. blog posts for hate groups, or enabling payment accounts for such actors. Working with these providers to identify such actors, and to provide friction in funding and disseminating such content may be a useful mechanism for countering online harms.	Allows verification / checks to be undertaken through existing technology e.g., card verification. Enables 'mainstream' web hosts and payment processors to jointly recognise and help tackle online harms (e.g., denying payments to hate groups) Restricts the ability of hate groups to monetise content May encourage a culture of self-regulation (i.e., removal by hosts of problematic content early)	May create uncertainty regarding domains / access to payments for certain groups e.g. potential for payment processors to be a 'choke-point' for online speech – and disenfranchising payments to certain groups such as sex workers. *"As long as businesses like OnlyFans are reliant on centralised tech infrastructure, they will always behave like businesses that are 'renting' & not 'owning,' and they'll always be scared that their*

				landlords (Mastercard/Visa, Paypal, Amazon Web Services) will evict them,"[186]
Advisory / Consultancy / Representative	e.g., AVPA, Island23	This refers to organisations advising or supporting 'safety-by-design' or informing or representing users about broader trust and online safety.	Improves awareness of trust and safety Improves education about online safety and how to respond to online harms Can lead to the implementation of technical and non-technical approaches to counter online harms	
Advocacy / Non-profit	e.g., GDI, Moonshot, IWF	This refers to organisations founded with a particular purpose to address (one or more) online harm(s). They often have a revenue stream either consisting of contracts or grants, but may also be structured similar to a social enterprise (not-for-profit / surplus reinvested into R&D etc)	Enables significant expertise to develop in addressing online harms in a commercial setting Can hold platforms to account (demonstrate how online harm can be addressed) and criticise inaction etc.	Can require external support / funding to sustain the business model May not have the commercial scale to fully realise particular challenges (e.g., costs of accessing social media data)

[186] Issie Lapowsky (2021). OnlyFans reveals Visa and MasterCard's hold on online speech. [online] Protocol — The people, power, and politics of tech. Available at: https://www.protocol.com/policy/onlyfans-visa-mastercard.

Further, we also consider that the market can be segmented by customer type (typically sector or use-case based). At a high level, these include:

Table 5:2 Market Segmentation by Customer Type

Customer Type / Sector	Subsectors	Example organisations:	Example approaches used
Gaming and Entertainment	Gaming, Video Streaming, Immersive Tech, Sports, Adult, Gambling	e.g., EA, Twitch, bet365, DFB	- Identify / age verification (age-gates and individual verification) - Content moderation to tackle toxicity - Age-appropriate design - Identification of harmful actors (in online / real-life) e.g., racial abuse online of footballers.
Media	Advertising, Marketing, TV / News Media, Print	e.g., BBC, Sky, WPP	- Use of fact-checking - Identification of false or misleading narratives - Identification of disinformation / harmful actors
Retail / Services	Retail, Food, Delivery, Accommodation, Toys	e.g., LEGO, Amazon, Airbnb	- Identification of illegal content (e.g., third party sale of extremist paraphernalia) - Use of age-appropriate design - Identity verification (for user safety) - Identification of disinformation / advertising links to harmful sites
Social Apps and Platforms	Social Media, Dating, Search Engines, Forums, Workplace	e.g., Match Group, Twitter,	- Identification of illegal and harmful material - Content moderation

	Safety	Google, Culture Shift	- Use of age-appropriate design - (Individual level) use of parental controls and filtering
Telecoms and Digital Infrastructure	ISPs, Telecoms, Data Centres, Payment Processors	e.g., Telefonica, VISA, Mastercard, Shopify	- Verification of transactions / identity - Age assurance and filtering (e.g., mobile access for under 18s)
Public Sector	Education, Health, Policing and Law Enforcement	e.g., filtering within *schools*, identifying terrorist content within *police forces*, flagging self-harm material etc.	- Web Filtering (education) - Digital Forensics - Social media / threat intelligence

However, each of these sector/customer types will have differing ease of access depending on the technology or solution being offered, and the current approach to online safety. For example, it can be challenging for new safety tech providers to generate commercial relationships with existing social media that already have developed their own in-house approaches (or require third parties to be at a certain scale).

This means that generating success stories or positive use-cases can be a good mechanism for safety tech companies to further their business model e.g., establishing partnerships with games developers, or football associations to tackle common problems of online harm.

Current levels of investment in trust and safety

It is difficult to identify the existing levels of internal expenditure on trust and safety (and safety tech) across current platforms given the privacy and commercial sensitivity within these. However, some platforms have disclosed high-level estimates for their investments in safety. These are explored below (and are estimates only based on identifiable content):

- **Facebook**: Invested $13bn in content moderation[187] and addressing misinformation on their platform since 2016, with 40,000 staff (assumed in-house and outsourced) working on safety issues. In context, Facebook's aggregate revenues between 2016-2020 were $281bn[188], suggesting expenditure on safety and security is approximately 4-5% of revenue. Further, with 40,000 people working in this space and a user community of c. 2.9bn – this suggests one 'safety and security' professional for every 72,500 users. However, further granularity or definition of 'safety and security' is not available.
- **Google:** It is estimated that **"**c.10,000 people scrutinise **YouTube** and other Google products"[189].
- **Twitter**: It is estimated that Twitter has a team of c. 1,500 moderators[190]. In 2019, they acquired Fabula AI to help use ML techniques to analyse platform behaviour and abuse.
- **Reddit:** Reddit has approximately 700 staff globally. It is estimated that c. 10% of its workforce are involved in content moderation[191]. However, the majority of content moderation by voluntary moderators of individual subreddits (i.e., mods). Microsoft estimates that there are approximately 91,563 unique mods on the platform (an average of five mods per subreddit).
- **Gaming Platforms:** Major gaming companies such as EA, Infinity Ward and Vale have all launched anti-toxicity campaigns, developing moderation tools for voice chat, and improving the ease and transparency of player-reporting systems. Player uptake of tools is limited, however, with an Anti-defamation League survey suggesting that less than half of the total respondents are using the established report process. This is due to a number of reasons, such as the

[187] mint. (2021*). Facebook says it has spent $13 bn on safety, security since 2016 US election*. [online] Available at: https://www.livemint.com/companies/news/facebook-says-it-has-spent-13-bn-on-safety-security-since-2016-us-election-11632231393584.html [Accessed 18 Oct. 2021].
[188] Statista. (2018*). Facebook: annual revenue 2018* | Statistic. [online] Available at: https://www.statista.com/statistics/268604/annual-revenue-of-facebook/.
[189] Barrett, P. (2020). *Who Moderates the Social Media Giants?* [online] NYU STERN Centre for Business and Human Rights. Available at: https://static1.squarespace.com/static/5b6df958f8370af3217d4178/t/5ed9854bf618c710cb55be98/1591313740497/NYU+Content+Moderation+Report_June+8+2020.pdf.
[190] ibid
[191] Singh, S. (2019*). Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content.* [online] New America. Available at: https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/case-study-reddit/.

effort required to submit a report, reports not being taken seriously, and the normalisation of toxicity within the gaming experience.[192]

5.5 Areas of Growth

Within the stakeholder consultations, and through a review of each of the safety tech sub-sectors, we note that despite its nascence, this is clearly a global marketplace ready for growth. However, despite typical metrics for market growth (e.g., increase in revenue, profitability, investment raised, or employment) being strong, there are some areas that appear to be particularly interesting from a future perspective.

Overall, with respect to addressing online hate, there are a number of markets that should be of interest. These include:

- **Identification of harmful actors:** To address online harms, it is essential to 'map the bad actors'. As one stakeholder consultee noted, one of the challenges in online hate is that currently many organisations see AI and content moderation in its own right as a 'solution'. However, it is important to take a step back, and consider the gaps in this approach, and to curate ethical datasets that focus on particular hate groups and their activities online.

  "Fund the identification of bad actors to generate high-quality data. Train the AI with high-quality data, and identify similar actors, super-sharers, and their behaviours and patterns"

  UK Safety Tech firm working on online hate and disinformation

- **Identification of online hate and disinformation:** In a similar light, being able to identify, define, and potentially rank or weight the impact of online hate and disinformation may be a useful tool in increasing awareness of particular hate actors and how to respond. For example, the work undertaken by organisations such as the Global Disinformation Index and NewsGuard to score content providers on their risk e.g., flagging alt-right 'news sources', and to share this information with platforms and advertisers may grow as awareness of the commercial risk of associating with such content grows.
- **Embedding trust and safety into the wider digital economy:** Despite the growth of the safety tech ecosystem, many platforms understand their own business model and customers better than anyone else. In this, encouraging established platforms to develop trust and safety teams, or to learn from others to implement new approaches (that work for their platform) should help to further the growth of trust and safety across the digital economy. For example, basic mechanisms such as being able to mute hateful accounts, quarantine harmful

[192] Wired. (2020). *Toxicity in Gaming Is Dangerous. Here's How to Stand Up to It*. [online] Available at: https://www.wired.com/story/toxicity-in-gaming-is-dangerous-heres-how-to-stand-up-to-it/.

content, or easily report hate and racist abuse if it occurs would all be beneficial stages if accepted as an industry standard.

- **Introducing real-world consequences for online hate:** There are arguably three main tools that can be supported by safety technology. These include introducing friction or removing users from sites where they breach terms and conditions, demonetising content, and use of law enforcement and response where appropriate. We expect that organisations that support the identification and response of such harms, with the potential to support real-world consequences (e.g., working with forensics or law enforcement) will be high-growth in future years.
- **Intervention based approaches:** As one stakeholder noted, we have so many types of online harm out there, which means we need to develop different methodologies and interventions to address these. The tech approach to responding to CSAM, for example, is profoundly different to the approach required to identify and respond to anti-Semitism.
- **Responding to online hate across platforms:** Where possible, solutions should allow for information sharing across platforms and platform types to help tackle online harms more systematically rather than in silos.

Within the consultations, content moderation was identified as potentially requiring innovation to meet some inherent challenges.

- **Content moderation:** The term 'content moderation' as discussed previously is inherently broad. It can include simple lexicon checks and manual reviews, that hold significant subjectivity where subject to AI techniques, and whilst it may require human input, this has a number of human and ethical considerations at play. Where new technology can reduce the burden on human moderators, or better detect hateful motivations or behaviours, this will be a useful ambition of content moderation start-ups. However, there is a risk that content moderation could be used as a 'sticking plaster' and must require continual innovation to improve current practices. Further, the risk of over-moderation or false positives may also negate some of the wider benefits of removing harmful content.

"*Many platforms are still lexicon-based when it comes to content moderation. This means they're taking content action, but not account action. This means that it's a very broad-brush effect, applied within a very narrow context."*

Stakeholder interview, Safety Tech SME working on online hate

"*People are getting caught up where there's no real issue, but the words they have used mean they're in the identifier." (as above)*

5.5.1 Investment in Safety Tech

Within the UK Safety Tech Sectoral Analysis (2020), this report highlights increasing investor interest in trust and safety solutions, reflected by increased volume and value of investment raised by safety tech businesses.

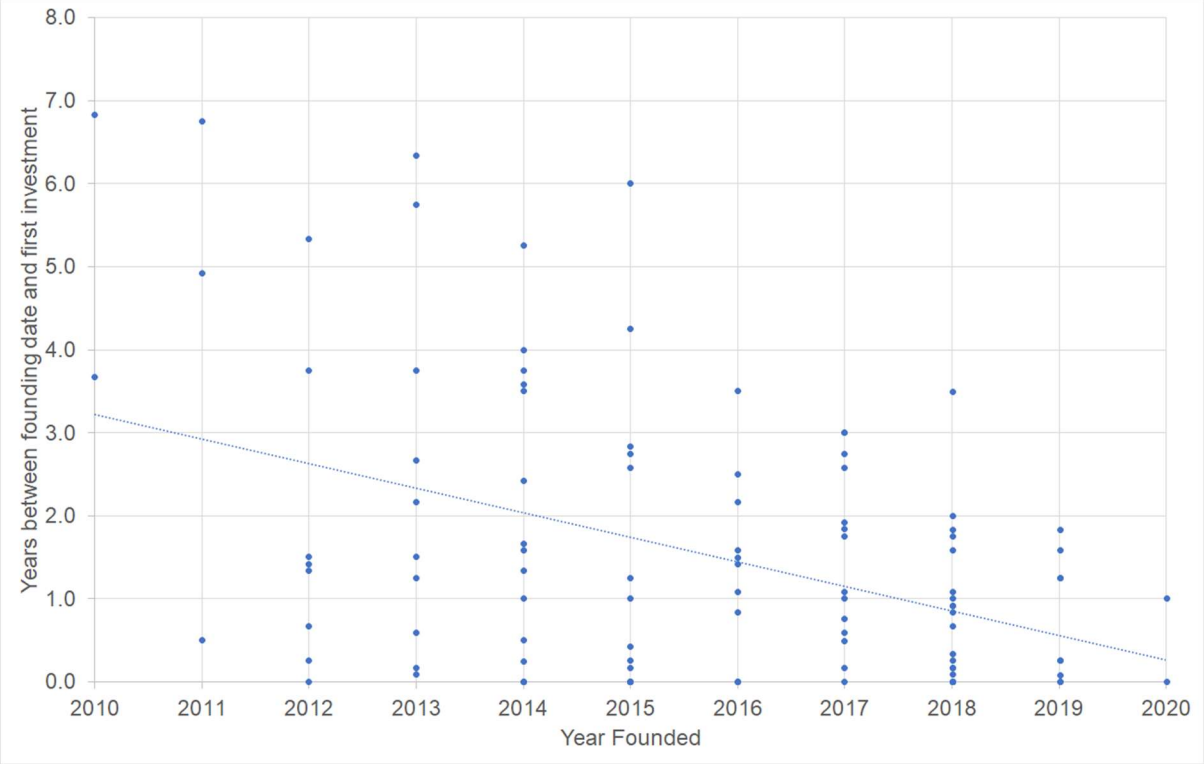
The Alfred Landecker Foundation is currently working with Dealroom to develop a database of safety tech companies that are high-growth or have received some stage of external investment.

The research team has identified 110 'dedicated' safety tech businesses globally with a known founding date (x-axis) in receipt of external investment. We have measured the time taken between each business' founding date and their first (typically seed-level) investment.

This suggests that the average time taken to secure an initial investment among safety tech businesses has fallen significantly in recent years. This suggests an increasing demand and a need for seed investment within the sector. Further, the encouragement of academic start-ups within this domain may also enable investors to back particularly early-stage / stealth start-ups.

Time(s) taken between company founding and first investment

Figure 5:1 Safety Tech investment overview



Source: PE analysis of ALF / Dealroom data

6 Theme 4: Legal, Political, and Ethical

This section provides an overview of the key legal, political, and ethical considerations of Safety Tech. An overview of key findings is outlined below and include:

- **Legal:** There is an emerging consensus across Western countries around what constitutes harm in the online environment. This is supported in part by EU law but may become hindered due to similar laws elsewhere that may use online harm legislation to limit free speech;
- **Political:** Online platforms are facing increased political pressure. There are also wider concerns around the future moderation of harm and the implications of an existing focus on Western/ English speaking nations;
- **Ethics:** As solutions become more complex there will be an increased need to ensure that they are explainable and replicable. This will help address transparency and reporting issues within the sector and for third-party applications. The need for ethical design and the consideration of ethics and wellbeing at each stage of product development.

6.1 Legal

The section below provides a high-level overview of the legal landscape of Online Harms globally. Recent legislation shows that across the EU and further afield work is ongoing to define the scope of legal responsibility and remit relating to harm online.

There is a clear political momentum that can support broader conversation and increased engagement in the area of online harm that can help define harm and distribute responsibility fairly between the individual, private firms, and the legislature. Clear examples of this include pressure placed on Facebook to change how they engage with subcontract content moderators, and the landmark Australian case placing the responsibility for toxic comments on the content publisher.

> *"It is critical to distinguish Safety Tech from cybersecurity, shifting the focus from IT to humans is important. It will take a while for the Safety Tech sector to mature; an umbrella of policies are required, specifically protocols around best practices in Safety Tech"* (Stakeholder, Senior Law Enforcement Officer)

6.1.1 Legal obligation to combat online harms

The table overleaf[193] provides an overview of the online harm legal landscape, detailing the type of businesses in scope under each law. It should be noted that similar laws have also been enacted in other countries such as Spain, Russia, Venezuela, the Philippines, Kenya among others. Table 6:1 Services legally required to address harm under national law

Services in scope	Social media platforms	Cloud hosting	Video sharing platforms	Video games with user interaction	Online marketplaces	Search engines	Private or user-to-user interaction
Germany (NetzDG)	✓	✓	✓	×	×	×	×
France (LCEN and Avia Law)	✓	✓	✓	✓	✓	✓	✓
UK (Online Safety Bill)	✓	✓	✓	✓	✓	✓	✓
Ireland (Online Safety and Media Regulation Bill)	✓	✓	✓	✓	✓	✓	✓
Australia (BSA, AVMA, EOSA)	✓	✓	✓	✓	✓	✓	✓
Singapore (Internet Code of Practice, POFMA)	✓	✓	✓	✓	✓	✓	✓
USA (Section 230)	✓	✓	✓	✓	✓	✓	×
EU (DSA)	✓	✓	✓	✓	✓	✓	unknown

Source: Linklaters

[193] Packer, B. (2021*). Online Harms: A comparative analysis. [*online] Linklaters. Available at: https://lpscdn.linklaters.com/-/media/digital-marketing-image-library/files/01_insights/thought-leadership/2021/april/online-harms---a-comparative-analysis.ashx?rev=1c44d739-086d-400a-8f94-508a23148e5e&extension=pdf&hash=63F3E4D64476F056E124CD70774B33A8.

A more in-depth look at some of the major markets is provided below:

Table 6:2 Legal response to online harm

Law	The legal response to online harm
National	
Germany: NetzDG Act[194]	● Came into force October 2017; ● Social media must remove manifestly unlawful content within 24-hours; ● Fine issued between €5m and €50m
France: Avia Bill[195,196]	● Draft published May 2020; ● Supports existing laws which issue takedown notices for content that provokes or glorifies terrorist acts or relates to child pornography; ● Avia law would modify existing laws decreasing the takedown timeframe to 24 hours; ● Fine issued between €250k and €1.25m; ● **Avia law was deemed unconstitutional by French Constitutional Council June 2021 on the decision it was "not necessary, appropriate and proportionate"**
United Kingdom: Online Harms White Paper[197]	● Sets out government plans for a world-leading package of online safety measures that support the digital economy; ● Proposes the establishment of an independent regulator to uphold new laws on the duty of care of online platforms covering both harmful and illegal content.
Ireland: Online Safety and	● Ireland is establishing a new Online Safety Commission to deal with harmful online content, working with platforms on standards and defined legal issues rather than as a helpline to the public;

[194] Spiegel, J. (2018). *Germany's Network Enforcement Act and its impact on social networks*. [online] TaylorWessing. Available at: https://www.taylorwessing.com/download/article-germany-nfa-impact-social.html.
[195] Schuler, M. and Znaty, B. (2020). *New law to fight online hate speech (Avia law) to reshape notice, take down and liability rules in France.* [online] TaylorWessing. Available at: https://www.taylorwessing.com/en/insights-and-events/insights/2020/05/new-law-to-fight-online-hate-speech-in-france [Accessed 18 Oct. 2021].
[196] European Digital Rights (EDRi). (2020*). French Avia law declared unconstitutional: what does this teach us at EU level?* [online] Available at: https://edri.org/our-work/french-avia-law-declared-unconstitutional-what-does-this-teach-us-at-eu-level/ [Accessed 18 Oct. 2021].
[197] Department for Digital, Culture, Media & Sport (2019). *Online Harms White Paper*. [online] GOV.UK. Available at: https://www.gov.uk/government/consultations/online-harms-white-paper.

Media Regulation Bill[198]	● Permitted to seek to hold influential position holders in a designated online service criminally liable, in cases where the designated online service fails to comply with a warning notice from the new Online Safety Commission; ● The Bill is an important piece of legislation, which will see the creation of robust regulation for online platforms, setting Ireland apart as one of the first countries in the world to do so in a systemic way.
Australia (BSA, AVMA, EOSA)	● Australia's Broadcasting Service Act (1992) has broad application, Schedules 5 and 7 focusing on content hosted in or outside Australia; ● The Abhorrent Violent Material Act (2019) applies to material hosted on a "carriage service" (telephone or internet services) which is provided within or outside Australia; ● The Enhancing Online Safety Act 2015 applies to cyberbullying relating to an Australian child provided or posted on a social media service.
Singapore (Internet Code of Practice, POFMA)	● Broadcasting (Class Licence) Notification regulates prohibited harmful content in Singapore. Those within scope are required to abide by the conditions in the Internet Class Licence and to ensure that content on their platforms complies with the Internet Code of Practice, introduced in October 2016 ● Protection from Online Falsehoods and Manipulation Act (2019) otherwise known as the "fake news law".
USA (Section 230)	● Section 230 of the Communications Decency Act of 1996 (the "CDA") provides that "no provider or user of an interactive computer service shall be treated as the publisher or speaker of any of the information provided by another information content provider"

[198] Richardson, Z. (2021*). Online Safety and Media Regulation Bill: Social media firms facing hefty fines and criminal liability if they fail to meet new online safety standards.* [online] Fieldfisher. Available at: https://www.fieldfisher.com/en-ie/locations/ireland/ireland-blog/online_safety_and_media_regulation_bill [Accessed 18 Oct. 2021].

	● This means that within the USA, no legislation requires platforms to take measures in respect of harmful content online.
International	
UN International Convention on the Elimination of all forms of Racial Discrimination (1968) | General Recommendations on Combating Racist Hate Speech (2013)[199]	● Recognises the use of indirect language to disguise hate speech, especially when attempting to appear moderate ● Call for the criminalisation for: o The spread of hate speech o Inciting others to hate o Threatening others in the context of hate or inciting others to do the same o Offensive hateful speech that is motivated by inciting others to hate o Membership of hate-related groups that incite hatred ● Suggested techniques to address online hate include legislation that governs the operation of social media and internet providers within State jurisdiction, drawing on international standards; ● Hold social media and Internet providers accountable and impress user responsibility for disseminating ideas and opinions; ● Adoption of professional ethics by social media and internet providers that incorporate respect for the principles of the Convention and other fundamental human rights standards
Council Framework Decision 2008/913/JHA[200]	● Mentions specific criminalisation of online speech that: o Incites racist or xenophobic hatred or violence made via information systems; o Condones, denies, or grossly trivialises crimes against humanity, war crimes, and genocide that also incites hatred or violence made via information systems; o Public distribution of pictures or other material via information systems in the commission of either of the above acts
European Union: Code of Conduct on Countering	● Agreement between Facebook, Microsoft, Twitter, YouTube, Snapchat, Dailymotion, Jeuxvideo.com, and LinkedIn and the European Commission; ● The agreement includes a regular monitoring exercise using an agreed methodology;

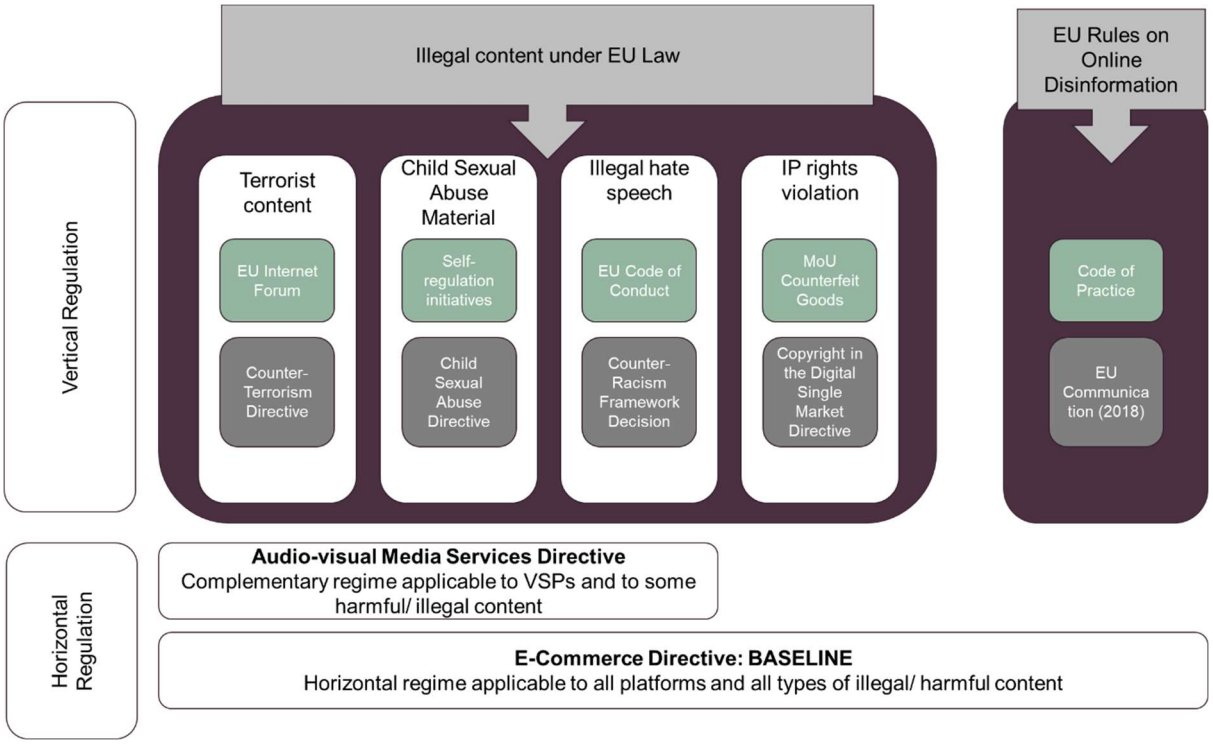
[199] Williams, M. (2019). *Hatred Behind the Screens A Report on the Rise of Online Hate Speech*. [online] Available at: https://hatelab.net/wp-content/uploads/2019/11/Hatred-Behind-the-Screens.pdf [Accessed 18 Oct. 2021].
[200] IBID

Illegal Hate Speech Online	● Social media firms now assessing 90% of flagged content within 24 hours and 71% of material deemed illegal is now removed
European Union: Digital Services Act Digital Markets Act[201]	● Aims to create a safer digital space in which the fundamental rights of all users of digital services are protected; and ● To establish a level playing field to foster innovation, growth, and competitiveness both in the European Single Market and globally; ● Focus on trade and exchange of illegal goods, services, and content online ● An effort to consolidate and harmonise country-level laws, e.g. Avia and NetzDG

[201] European Commission. (2021). *The Digital Services Act package | Shaping Europe's digital future*. [online] Available at: https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package.

The various horizontal and vertical regulations that exist under EU law specific to content moderation are also presented below. This shows that some stricter rules apply to video-sharing platforms and certain types of illegal harm, and vertical rules which are specific to particular content illegal under EU law (terrorist content, child sexual abuse material, racist and xenophobic hate speech, and violations of Intellectual Property).

Figure 6:1 EU Regulatory framework for online content moderation[202]



Source: Online Platforms' Moderation of Illegal Content Online Law, Practices and Options for Reform

[202] DeStreel, A., Defreyne, E., Jacquemin, H., Ledger, M. and Michel, A. (2020). *Online Platforms' Moderation of Illegal Content Online Law, Practices and Options for Reform*. [online] Available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/652718/IPOL_STU(2020)652718_EN.pdf.

6.1.2 Which countries or institutions are setting international norms in Safety Tech?

The future role of the Safety Tech sector was discussed in detail at the UN Internet Governance Forum in November 2020.[203] At the forum, online safety was flagged as a global issue, and it was noted that the sector is strategically placed to have significant economic and societal impacts in the future. Activity specific to individual countries is set out below:

United Kingdom The UK's Online Harms White Paper has led to serious conversations around online safety and the growth of an ecosystem in which Safety Tech can thrive. Swiss firm Privately is testimony to the growing UK ecosystem, having gained traction and support primarily within the UK market.

The UK's Online Safety Tech Industry Association (OSTIA) highlights the united vision between the UK Safety Tech sector and regulatory bodies, citing the alignment between Ofcom and OSTIA goals which has supported the sector's development in the country.

Platforms operating in the UK will have to meet a duty of care which will require them to: conduct risk assessments and ensure appropriate system and process is in place to protect against illegal and harmful content and activity.

Germany The German government is working in partnership with the UK on a series of #TechforGood talks, focusing on how technology can mitigate harm, and supporting collaboration between nations. Talks were held between the UK's Department for International Trade and the Department for Digital, Culture, Media and Sports, and the North-Rhine Westphalian Cyber Crime Agency ZAC.

Germany is noted as having increased media awareness domestically and is world-leading on legislation, but DIT trade advisors state that there are "some great products and services that Germany are currently lacking" that are in use in the UK.[204]

Germany is also addressing wider aspects of the platform landscape, and in January 2021 the German parliament agreed to reform competition law, with plans to introduce preventative measures to counter the market power of large digital platforms.[205]

Under the NetzDG providers operating in Germany will be required to implement a system for managing user reports and facilitating the reporting process. As mentioned, harmful content must be removed and reported to the Federal Criminal Office.

---

203 Francis, G. (2020). *Safety tech at the U*N. [online] www.safetytechnetwork.org.uk. Available at: https://www.safetytechnetwork.org.uk/articles/safety-tech-at-the-un [Accessed 18 Oct. 2021].
204 Safety Tech Innovation Network. (2021). *German safety tech industry gains momentum*. [online] Available at: https://www.safetytechnetwork.org.uk/articles/german-safety-tech-industry-gains-momentum [Accessed 18 Oct. 2021].
205 Van Dorpe, S. (2021). *Germany shows EU the way in curbing Big Tech*. [online] POLITICO. Available at: https://www.politico.eu/article/germany-shows-eu-the-way-in-curbing-big-tech/ [Accessed 18 Oct. 2021].

Ireland In December 2020 the Irish government announced new provisions in the finalised general scheme of the Online Safety and Media Regulation Bill, which establishes a new Media Commission to replace the Broadcasting Authority of Ireland, which deals with video regulation of video-sharing platforms, including YouTube, targeting criminal content and online harms.

This legislation is in line with the EU's Audiovisual Media Services Directive (2018) and will lead to the establishment of Ireland's Online Safety Commission to tackle online harms. This legislation will support the creation of robust legislation in Ireland, which is vital given that the country is home to several headquarters of the larger tech firms.[206]

There are also wider pressures from the Irish legislature's (Oireachtas Éireann) media committee who claim that social media firms need to do more to address harms online.[207] As outlined previously there are also concerns around the use of human moderation for training AI, which has been raised to the Irish parliament's Enterprise, Trade and Employment committee and may lead to changes in laws around outsourced human moderation teams and the role of NDA agreements within contracts.[208]

Similar to both the UK and Germany firms must have a mechanism in place to handle user complaints and mechanisms and carry out risk impact assessments on their platforms, while also having reporting obligations that must comply with Online Safety Codes.

Australia Australia has passed legislation that forces social networks to remove harmful content within 24 hours. Australia's Online Safety Bill will also require companies to provide identity and contact information about abusers on their platform. The bill was developed in response to the Christchurch incident in New Zealand and gives Australia's eSafety Commissioner powers to rapidly block websites.

The bill also strengthens existing penalties for online abuse and harassment, including up to five years imprisonment, and will require companies to keep an updated Online Content Scheme to do more to keep users safe online and give the eSafety Commissioner powers to require app stores to remove products enabling the provision of harmful kinds of online content.

---

[206] Richardson, Z. (2021). *Online Safety and Media Regulation Bill: Social media firms facing hefty fines and criminal liability if they fail to meet new online safety standards.* [online] Fieldfisher. Available at: https://www.fieldfisher.com/en-ie/locations/ireland/ireland-blog/online_safety_and_media_regulation_bill [Accessed 18 Oct. 2021].
[207] Slattery, L. (2021*). "Wild West" social media firms criticised for response to harmful content.* [online] The Irish Times. Available at: https://www.irishtimes.com/business/media-and-marketing/wild-west-social-media-firms-criticised-for-response-to-harmful-content-1.4569648 [Accessed 18 Oct. 2021].
[208] Nast, C. (2021*). Facebook's content moderators are fighting back.* [online] Wired UK. Available at: https://www.wired.co.uk/article/facebook-content-moderators-ireland [Accessed 18 Oct. 2021].

The bill has raised a number of concerns, around potential overreach, and impact on free speech. Twitter also raised concerns around impacts on smaller operations and the scope of defined harm.[209]

In addition to wider platform legislation, a defamation case from 2020 has the potential to impact the scope of responsibility held by platforms and media outlets, after courts ruled in favour of Dylan Voller who claimed that media outlets were responsible for "publishing" comments made under articles on their sites defaming his character. This opens wider doors around media companies' liability in defamation cases.[210]

The government is also holding social media firms accountable for algorithm preferences and requiring them to share details around news-generated revenue, and the sharing of data, ranking, and display of news content. This is all in an effort to ensure tech giants don't damage the market and competition.[211]

6.1.3 Implications for the Alfred Landecker Foundation

The above section highlights four core areas of legal concern that the Alfred Landecker Foundation should be cognisant of when supporting online harm mitigation. Outlined below these include:

- **Moderation vs. Freedom of Speech debate:** Spain's Citizen Security Law[212] is a key example of how ambiguous hate speech definitions can allow governments to limit freedom of expression and even lead to incarceration based on views stated online. It is important that online harm is accurately defined and adhered to. A concise definition is vital for lawmakers and when developing ethical solutions to combat harm. The Alfred Landecker foundation should consider to what extent solutions funded address online harms, while also considering potential scope creep and detrimental impacts that limit freedom of speech.
- **Responsibility to protect users, responsibility to protect moderators:** Landmark engagement with sub-contract moderators in Ireland highlights that there is a greater need to support ethical training of artificial intelligence systems and to support transparency in current training practices. There is a potential opportunity to look at TikTok's model in Ireland, with TikTok recruiting staff previously sub-contracted to Facebook, or through the use of synthetic data sources.

---

[209]Scott, J. and Savov, V. (2021*). Australian Law Could Force Facebook, Google to Strip Content*. Bloomberg.com. [online] 23 Jun. Available at: https://www.bloomberg.com/news/articles/2021-06-23/australia-s-online-safety-bill-forces-platforms-to-strip-content.

[210]Douglas, M. (2020). *Media companies can now be held responsible for your dodgy comments on social media*. [online] The Conversation. Available at: https://theconversation.com/media-companies-can-now-be-held-responsible-for-your-dodgy-comments-on-social-media-139775.

[211]Yano, A. (2020). *The Australian government holds Facebook and Google accountable.* [online] The American Genius. Available at: https://theamericangenius.com/social-media/the-australian-government-holds-facebook-and-google-accountable/ [Accessed 18 Oct. 2021].

[212] Matjašič, P. (2021*). Spanish gag law: The original sin and ongoing penance*. [online] Al Jazeera. Available at: https://www.aljazeera.com/opinions/2021/3/1/spanish-gag-law-the-original-sin-and-ongoing-penance.

- **Legal responsibility of "publisher":** Dylan Voller's defamation case in Australia offers potential insight into the direction of future legislation and the responsibility of harmful content online. In Voller's case, responsibility was given to the media outlet that produced the initial content, who offered the platform and promoted engagement from users. If this precedent continues or is mirrored elsewhere news outlets may be held accountable for publishing politically charged or misleading headlines that lead to harmful engagement online. Engagement with news outlets to moderate their own comment sections offers a potential avenue for intervention that does not rely on AI-driven or mass moderation by social media firms.
- **Supporting a unified, multinational approach**: While there is an emerging consensus across jurisdictions and measures put in place to reduce harms, there is also a need within the Safety Tech sector that calls for a strategically placed and unified approach when addressing online harm. With this being said, it will be important that the Alfred Landecker Foundation engages with the sector, and government across different jurisdictions to ensure funded interventions do not duplicate existing efforts, or do not undermine ambitions due to scope creep (i.e., inhibiting freedom of speech, or other links to controversy).

6.2 Political / International Context

6.2.1 The consequence of moderation on freedom of speech

Germany's approach to online harm has been mirrored by at least 13 countries globally. Of these countries, four have been ranked as "not free" (Venezuela, Vietnam, Russia and Belarus, and Honduras); five are ranked "partly free" (Kenya, India, Singapore, Malaysia, and the Philippines) and only three are "free" (France, the UK, Australia).[213]

This raises concerns about how changing norms in online moderation can impact freedom of speech across the globe, in light of a widened scope of "harmful content" that has been listed to include "fake news," "defamation of religions," and "anti-government propaganda".

Of the above countries, Vietnam is the most explicit, prohibiting "propaganda against the Socialist Republic of Vietnam,". Like Russia, Vietnam has set up a government body that will provide takedown notices and that is not open to independent review.

Having said this, countries like the UK and India have incorporated a "duty of care" into their law, requiring intermediaries to actively police and preventively remove illegal or undesirable content.

The key distinction in norms across these regions is the extent to which each country incorporates freedom of speech or is open to public scrutiny.

In countries with increased censorship, access to an online space can be fundamental in supporting political change. Vietnam is a key example, with Amnesty International noting how the online environment has supported political discourse safely, allowing activist group Viet Tan, which is illegal in Vietnam, to work remotely from the US.[214]

[213]Fiss, J.M., Joelle (2019). *Germany's Online Crackdowns Inspire the World's Dictators*. [online] Foreign Policy. Available at: https://foreignpolicy.com/2019/11/06/germany-online-crackdowns-inspired-the-worlds-dictators-russia-venezuela-india/ [Accessed 18 Oct. 2021].

[214] Logan, S. (2018*). Facebook and Vietnam's new cybersecurity law*. [online] The Interpreter. Available at: https://www.lowyinstitute.org/the-interpreter/facebook-and-vietnams-new-cybersecurity-law [Accessed 18 Oct. 2021].

6.2.2 Language and context

Another key consideration for Safety Tech in the global context is the extent to which the sector can support minority languages.

As an example, more than 90% of Facebook's monthly users are non-English speaking, despite the majority of the platform's moderation budget supporting English language-based AI. This has raised concerns over how effective different tech companies are at monitoring harms outside of English-speaking nations.[215]

In Facebook's case, former employees have stated that internal activity to address this issue is minimal and considered a "cost of doing business".

In the case of AI-moderation, language can be considered a resource, and there is a lack of available resource in the Global South,[216] which makes research and application of AI models harder to accomplish.

Work however is undertaken by researchers focused on minority groups, e.g., Wijeratne's language corpora supporting analysis for the colloquial Sinhala community in Sri Lanka.

Engagement or facilitation of smaller research teams may be able to address this problem or remove the bias associated with English language models and research teams.

> *"The top challenge centres on providing Safety Tech solutions to the Global South - Internet use has risen rapidly amongst children and young people in this area. Online harm and online safety are global challenges - we, therefore, need to implement global solutions"*
>
> (Academic stakeholder who conducts research in the Global South)

[215]Canales, K. (2021). *Facebook's AI moderation reportedly can't interpret many languages, leaving users in some countries more susceptible to harmful posts*. [online] Business Insider. Available at: https://www.businessinsider.com/facebook-content-moderation-ai-cant-speak-all-languages-2021-9?r=US&IR=T [Accessed 18 Oct. 2021].
[216]Wijeratne, Y. (2020). *Facebook, language and the difficulty of moderating hate speech* | Media@LSE. [online] London School of Economics and Political Science. Available at: https://blogs.lse.ac.uk/medialse/2020/07/23/facebook-language-and-the-difficulty-of-moderating-hate-speech/ [Accessed 18 Oct. 2021].

6.2.3 Privacy Considerations

The various online harms bills included above show the different approaches that countries are taking to address harm and the different definitions used in their approach to online harm. These definitions are central to the privacy debate which is biased towards a western perspective.[217]

"Knowledge and learnings from the so-called global south are still treated as one-off case studies when they should be contributing towards… more nuanced and inclusive conversations" around what privacy means globally.

There is a growing body of academics who are viewing digital privacy from a more intercultural perspective, assessing what privacy means across cultures. This is vital considering between China and India there are c.2.8bn inhabitants. They believe that digital ethics and privacy should be viewed in the local context and based on ground realities and cultural legacies, and to ensure that the politics that underpin the design and structure of technology and systems can be held to scrutiny.

There is growing pressure on tech firms to provide access to data to support crime detection. Some governments believe that such backdoors can be targeted to provide access to public authorities only, so they can access the digital evidence they seek to prosecute criminals using encrypted communications.[218]

Having said this, various scandals offer insight into the insidious nature of surveillance, e.g., the Cambridge Analytica scandal and Clearview AI's facial recognition system. More recently COVID-19 track and trace data in Singapore is now being used by law enforcement, highlighting the dangers of large data sources and the wider debate around data ownership. While there is the debate that access and monitoring of encrypted data can prevent crime there are also risks that identification technology will be used to target minority groups. A recent high-profile example includes the use of Huawei AI software to identify Uighur minority group members in China.[219]

> *"We need to be cognisant of the tension between safety, security, privacy and online freedoms - particularly in the USA. Smart Safety Tech solutions could solve for that tension in a global context"*
>
> (Stakeholder, venture capital)

[217] Venkataramakrishnan, S. (2021). *Online privacy: a fraught philosophical debate.* [online] www.ft.com. Available at: https://www.ft.com/content/50441ea2-bb0b-4b10-90b9-941ffd262f82.

[218] European Internet Forum. (2021). *European Internet Forum - Decrypting the encryption debate: How to ensure public safety with a privacy-preserving and secure Internet?* [online] Available at: https://www.internetforum.eu/events/events/1127-decrypting-the-encryption-debate-how-to-ensure-public-safety-with-a-privacy-preserving-and-secure-internet.html [Accessed 18 Oct. 2021].

[219] Harwell, D. and Dou, E. (2020*). Huawei tested AI software that could recognize Uighur minorities and alert police, report says. Washington Post*. [online] 8 Dec. Available at: https://www.washingtonpost.com/technology/2020/12/08/huawei-tested-ai-software-that-could-recognize-uighur-minorities-alert-police-report-says/.

6.3 Ethics

6.3.1 Ethical governance and design

As previously stated, AI-driven harm prevention is trained using sample data from online platforms. To develop effective solutions service providers must therefore assess and train AI using accurate data sources. Access to these data sources has been cited as one of the key barriers to innovation within the sector.[220]

To support solution development the UK government has announced £2.6m worth of funding as part of their National Data Strategy to support innovation and competition within the Safety Tech Sector in an ethical way.

This £2.6m will go towards improving classification and sharing of data to support the detection of online harms such as cyberbullying, harassment, or suicide ideation.[221] It will involve the review and upgrade of government data standards and consultation across the public and private sectors and has been informed to date by the Government Office for Science's The Future of Citizen Data Systems report.[222]

Examples of unintentional or problematic use of big data are widespread and include Los Angeles's case against IBM and their misuse of data collected through its weather app; unintentional bias within Optum services which allegedly recommend better treatments to white patients; as well as Facebook's collaboration with Cambridge Analytica which shared personal data of more than 50m users.

With the side effects of poor model design or unethical use of data becoming more pronounced, larger tech firms are now building out ethical teams and operationalising data and AI ethics.

Key recommendations for ethical data design at a business level are summarised below[223]:

- **Identify existing governance structure that a data and AI ethics program can leverage:** Incorporate ethics into wider discussions around cyber, risk, privacy, and analytics.
- **Create a data and AI ethical risk framework that is tailored to the industry:** This should be informed by engagement with stakeholders and articulate how the

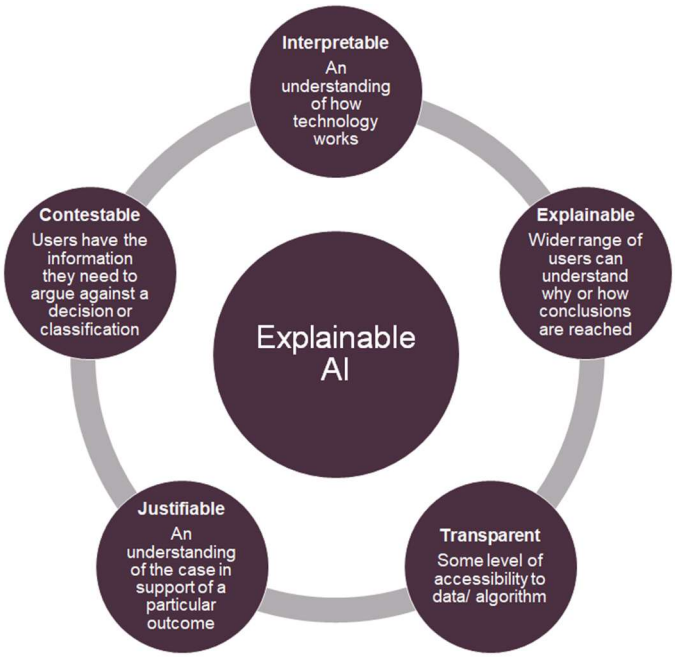
[220]Francis, G. (2021). *New industry consortium launches to transform access to online harms data*. [online] Safety Tech Network. Available at: https://www.safetytechnetwork.org.uk/articles/new-industry-consortium-launches-to-transform-access-to-online-harms-data [Accessed 18 Oct. 2021].
[221]Department for Digital, Culture, Media & Sport. (2020). *Government publishes new strategy to kickstart data revolution across the UK*. [online] Available at: https://www.gov.uk/government/news/government-publishes-new-strategy-to-kickstart-data-revolution-across-the-uk [Accessed 18 Oct. 2021].
[222] Government Office for Science. (2020). *The future of citizen data systems*. [online] Available at: https://www.gov.uk/government/publications/the-future-of-citizen-data-systems.
[223] Blackman, R. (2020*). A Practical Guide to Building Ethical AI. Harvard Business Review.* [online] 15 Oct. Available at: https://hbr.org/2020/10/a-practical-guide-to-building-ethical-ai.

ethical design will be maintained with future developments. This should include quality assurance, KPIs, and log and mitigation of risks.

- **Use healthcare ethics for insight**: Ethics have been forefront in healthcare since the 1970s and key concerns such as privacy, self-determination and informed consent have been addressed.
- **Optimise guidance and tools for product managers:** Customised tools should be developed that allow AI decisions to be both accurate and explainable. Work should be undertaken to promote the design and uptake of **explainable AI,** potentially supporting the incorporation of explainable AI into policy or design regulation.
- **Rewarding ethical decision making:** Promote the incentivisation of, and incorporation of ethics as a company's values.

Further insight from the UK House of Lords AI Committee, the EU's High-Level Group on AI, and the USA's Defence Advanced Research Projects Agency provides an overview of what explainable AI looks like:

Figure 6:2 Explainable AI Overview



Source: UK House of Lords AI Committee

Peters (2019)[224] also provides an overview of the ethical design process through their Responsible Design Process for Tech, which they developed in line with feedback

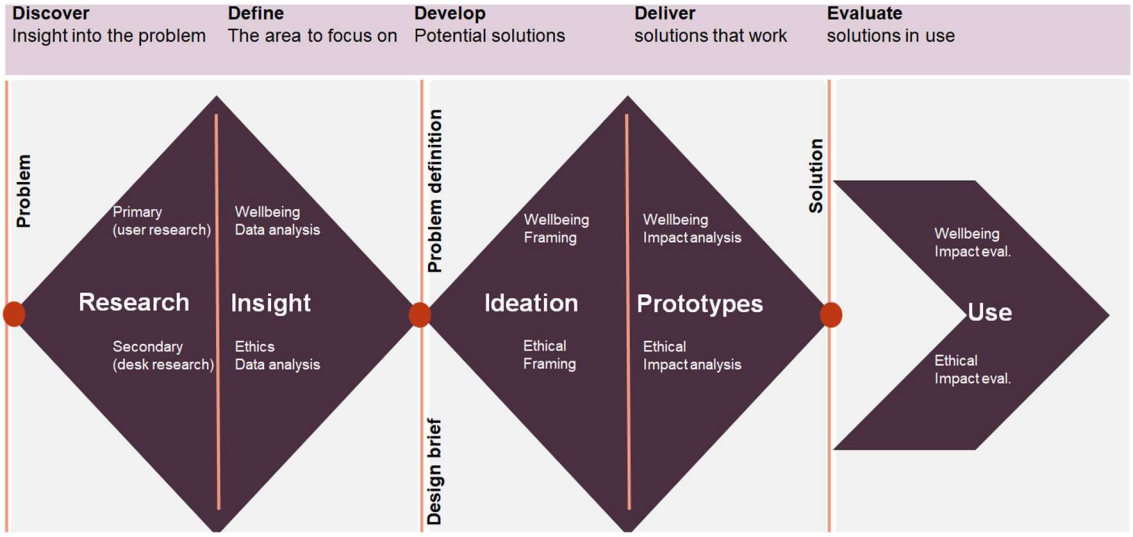
[224] Calvo, R. and Peters, D. (n.d.). *About*. [online] Responsible Tech Design. Available at: https://www.responsibletechdesign.com/p/blog-page.html [Accessed 18 Oct. 2021].

from colleagues at the Leverhulme Centre for the Future of Intelligence at the University of Cambridge[225]:

Key stages of the process are outlined in detail below:

- **Research:** Research of the needs, preferences, contexts, and lives of the people who use or are served by the technology. This can include secondary desk research and standard qualitative approaches (e.g., design thinking methods, ethnographic, workshops, and stakeholder engagement);
- **Insight:** Analysis of user research data through the lens of the wellbeing theory, identifying potential harms to wellbeing; and ethics data analysis through the lens of an ethical framework, identifying potential biases, ethical risks, or tensions.
- **Ideation:** Wellbeing framing involves incorporating wellbeing psychology concepts to help the design team determine the root psychological causes of user needs and to brainstorm solutions that are tailored to support digital wellbeing. Ethical framing makes the design team aware of ethical tensions.
- **Prototyping:** Wellbeing impact analysis involves collaborative speculation on the wellbeing impacts (good and bad) to which a particular design concept may lead. This should involve a wide range of stakeholders, including, of course, end-users. Ethical impact analysis involves collaborative speculation on the ethical impacts to which a particular design concept may lead.

Figure 6:3 Responsible Design Process for Tech



Source: Peters (2019)

[225]Peters, D. (2019). *Beyond principles: A Process for Responsible Tech*. [online] The Ethics of Digital Experience. Available at: https://medium.com/ethics-of-digital-experience/beyond-principles-a-process-for-responsible-tech-aefc921f7317 [Accessed 18 Oct. 2021].

Wider recommendations to support transparency in tech are outlined in the All-party Parliamentary Group's Trust, Transparency and Tech (2019) report.[226] This report includes 7 recommendations, outlined in full in the appendix with a focus on building public confidence, providing information to the public, supporting accountability, consistency, and increasing transparency around consent and use of AI.

6.3.2 Ethical data collection and consent

KPMG's 2021 survey into corporate responsibility found that 29% of director-level employees admitted to sometimes conducting unethical data collection. It has been noted though that the current legal landscape (e.g., GDPR) is influencing data collection processes with companies moving further away from a collect-all approach towards an outcome-driven approach to collection. Having said this, 70% of surveyed firms reported that they are still increasing the amount of personal information they are collecting.[227] Emerging solutions that support ethical AI and machine-learning are outlined below:

- **Use of synthetic data:** Synthetic data is cheap to produce and can support AI model development without exposing personally identifiable data. It is estimated that by 2024 that 60% of the data used to train AI will be synthetic. A 2017 pilot study also revealed that in 70% of cases synthetic data performed on par with real data, and the various advantages of its use include enhanced privacy, simulation of potential datasets, ability to overcome restrictions of real data, and an ability to focus on relationships. Disadvantages include the potential to miss outliers and the need for quality input data to produce the synthetic equivalent. Other challenges include the resource required to produce the data and to assess its output.[228]
- **Encrypted data:** This process involved securing and anonymising personal identifiable data prior to data training and modelling. As with synthetic data, however, the input of the data must be unbiased and accurate to real-world scenarios.[229]
- **Third-party inspectability:** This refers to allowing a third-party to examine systems to ensure they meet defined standards and decision making and is a core aspect of transparency. Inspectability is linked with "Explainability" above and relates to how traceable, verifiable, intelligent, and honest a design is.[230]

---

[226]Tindale, J. and Muirhead, O. (2019). *Trust, Transparency and Technology: Building Data Policies for the Public Good*. [online] Policy Connect. Available at: https://www.policyconnect.org.uk/research/trust-transparency-and-technology-building-data-policies-public-good [Accessed 18 Oct. 2021].
[227] Kahn, J. (2021). *Be afraid: Executives warn about personal data harvesting and use*. [online] Fortune. Available at: https://fortune.com/2021/08/24/eye-on-a-i-data-privacy-unethical-kpmg-survey/.
[228] Dilmegani, C. (2021). *The Ultimate Guide to Synthetic Data in 2020*. [online] AI Multiple. Available at: https://research.aimultiple.com/synthetic-data/.
[229] Priya, B. (2021). *Private AI: Machine Learning on Encrypted Data*. [online] OpenMined Blog. Available at: https://blog.openmined.org/private-ai-machine-learning-on-encrypted-data/ [Accessed 18 Oct. 2021].
[230] Felzmann, H., Fosch-Villaronga, E., Lutz, C. and Tamò-Larrieux, A. (2020). *Towards Transparency by Design for Artificial Intelligence*. Science and Engineering Ethics, 26(6), pp.3333–3361.

7 Key Findings and Discussion

This literature review has explored four key themes with respect to the incidence, pathways, and responses to online hate, and online harm. Through this research, we set out some key findings and discussion points that should guide further research, discussion, and action in this area.

7.1 Behavioural Science

1 **Working towards a definition of online hate:** Our research finds that there is 'no individual, globally recognised definition of online hate speech. There are, however, working definitions, as well as defined 'elements' of online hate. Further examination towards a shared or coherent definition of online hate, and its affected groups, would be welcome.

2 **Recognising the complexity of online hate and human behaviour online:** This literature review has identified several factors to consider with respect to human behaviour online, and its impact in generating and sustainable hateful content and conduct. The consideration of online hate identified several types of hate, actors, mediums, and responses - and this highlights the significant complexity and need for many approaches to identify varying strands of online hate. As one respondent within the stakeholder consultations noted, *"the response to tackling misogyny can be different to that of racism or to that of anti-Semitism".* Harmful online behaviour can be in part explained through pre-existing social science models. Integrating social scientists into design teams may support effective or intelligent design and embed 'safety by design' principles at an early stage.

3 The '**pathway to online hate' is not linear - it is multifactorial**. These factors are often developmental, cultural, environmental, event-influenced, platform-assisted, and algorithmic. It is therefore important to explore this further, to help develop a further understanding of root cause factors. Additionally, further understanding of the pathways towards online hate will enable the provision of specific interventions that can help address this.

4 **Learning from emerging behavioural responses to online hate:** We are still at an early stage, but there are solutions and methodologies available focused on the behavioural element of online hate such as counter-narratives and introducing friction. The literature review explores these and suggests that the stage these are implemented matters for the proliferation of online hate.

5 There is, however, more **limited evidence available on the efficacy of early-stage interventions** such as critical thinking and education in the information environment. This suggests both a market and public problem and should potentially reflect an area for the Alfred Landecker Foundation to invest or support exploratory work.

7.2 Technological

6 **Safety Tech approaches can be deployed as a barrier to hate:** Removing, blocking content, and de-platforming are often cited as leading to a challenging scenario as groups use alternative platforms. However, this research considers that the aim of 'safety tech' should be to remove such content or behaviour from mainstream platforms and create sufficient friction for alternative platforms to **diminish and suppress** the prevalence of hateful content.

7 **Online hate intelligence and identification of 'bad actors' appears to be a useful technological tool in identifying particularly harmful groups**, identifying members and prevalence of content, and assisting with takedowns, supporting friction, and notifying law enforcement and real-world sanctions.

8 **Establishing common international standards and data-sharing may be helpful to counter online hate**. For example, an open-source list of up-to-date hateful words, imagery, content, and use and context could be used to help disrupt the sale of extremist paraphernalia or disinformation on online marketplaces.

7.3 Economics

9 **Incentives matter across the entire online ecosystem.** We need to develop a better understanding of how platforms and individuals can benefit (e.g., financial, reputational, political, and otherwise) from sharing online hate and disinformation - and reduce this activity either through regulation, defunding, or removal for breach of terms. This literature review suggests that identifying such activity, and demonetising can be an effective tool. As mentioned previously, improved threat intelligence and knowledge sharing are critical components in disincentivising hateful content and behaviours online.

10 **Asymmetric information enables online harm**. Supporting knowledge sharing and process transparency is needed to understand harms on a cross-platform basis.

11 **There remains clear demand for investment in early-stage Safety Tech companies:** the investment landscape has shortened the time between launch and funding rounds. Start-ups are seeking out early-stage investment. This should position the Alfred Landecker Foundation in a place to offer seed / early-stage investments, and or consider co-investment with a larger partner. There may also be scope to engage with academic spin-outs whereby there is a particular focus on safety tech with respect to democratic values, protecting privacy and safety, or tackling online hate and disinformation.

7.4 Legal, Political and Ethical

12 There does appear to be some signs of **emerging legal cohesiveness** between states in addressing online harms. For example, the literature review highlights similarities and learning between countries such as Germany, Australia, the UK, and the EU. This is important as democracies explore what works and doesn't work well in online harms regulation.

13 **There will continue to be a 'safety vs privacy' debate. However, these do not have to be mutually exclusive terms**. For example, solutions that protect an individual's personally identifiable information may help to reduce online harms such as doxing and impersonation, as well as economic harms. There is the risk, however, with some 'safety tech' solutions (particularly with respect to identity or tracking individuals or groups with the perceived capacity to harm), that these could face challenges with respect to rights to privacy. It is therefore crucial that such tensions are explored, and solutions should demonstrate a respect for privacy, personal data, and all relevant legal requirements.

14 Further, the proliferation of online and social platforms means that there is a significant volume of data that exists regarding individuals and their behaviours and activities online. For example, this can be used for marketing, but it also runs the risk that personal data is used to estimate or link towards harmful or deviant behaviour in a problematic way e.g., using OSINT for risk scoring, which has a number of ethical considerations. **ALF should ensure that any investments or support provided is contingent upon an ethical framework for data curation and management and is GDPR complicit as a minimum.**

Appendix

Theme 1: Behavioural Science

Table A:1 Alt-tech platform overview

	Name	Description
Social media/ microblogging	Gab (2016)	Gab is an American alt-tech social networking service known for its far-right user base.
	Parler (2018)	Parler is an American microblogging and social networking service. It has a significant user base of Donald Trump supporters, conservatives, conspiracy theorists, and far-right extremists.
	MeWe (2012)	MeWe's light approach to content moderation has made it popular among American conservatives, conspiracy theorists, and anti-vaxxers.
	Minds (2015)	Minds is an alt-tech blockchain-based social network. Users can earn money or cryptocurrency for using Minds, and tokens can be used to boost their posts or crowdfund other users. Minds has been described as more privacy-focused than mainstream social media networks.
	Thinkspot (2019)	Thinkspot is an online social networking service similar to Patreon centered on free speech.
	WrongThink (2016)	Facebook-style social network.
	Substack (2017)	Substack is an American online platform that provides publishing, payment, analytics, and design infrastructure to support subscription newsletters.
Online video platform	BitChute (2017)	BitChute is a video hosting service known for accommodating far-right individuals and conspiracy theorists and hosting hate speech.
	DLive (2017)	DLive is an American video live streaming service that was founded in 2017. BitTorrent purchased it in 2019before
	DTube (2016)	Decentralised Tube (or DTube for short) is a YouTube-like video platform.

	Odysee (LBRY) (2015)	LBRY is a blockchain-based file-sharing and payment network that powers decentralised platforms, primarily social networks and video platforms. LBRY's creators also run Odysee, a video-sharing website that uses the network.
	PewTube (2017)	A YouTube-style video platform.
	Rumble (2013)	A YouTube alternative that has attracted an Alt-user base from August 2020.
Crowdfunding	GoyFundMe (2010)	Patreon-style funding platform.
	Hatreon (2017)	Patreon-style funding platform.
	SubscribeStar (2017)	Patreon-style funding platform.
	WeSearchr (2015)	Patreon-style funding platform.
News aggregator	Patriots.win (2015)	Reddit-style platform
	Voat (2014)	Reddit-style platform
Wiki encyclopedia	Infogalactic (2017)	Wikipedia-style platform
Imageboard	4chan (2003)	Imageboard platform containing alt-right content and Incel content.
	8chan (2013)	Imageboard platform containing alt-right content and Incel content.
	Slug	Imageboard platform containing alt-right content and Incel content.

Instant messaging	Signal (2014)	Messenger-style app
	Telegram (2013)	Messenger-style app
Online dating	WASP Love (2016)	Dating site for white nationalists and Christians.
Pastebin	JustPaste.it (2009)	Justpaste.it is a site that allows users to paste text and distribute the resulting link. The site became the object of international attention after supporters of the Islamic State began to use the site to disseminate information.
Domain name	Epik (2009)	Epik is an American domain registrar and web hosting company known for providing services to websites that host far-right, neo-Nazi, and other extremist content.
Civic engagement	CloutHub (2018)	Originally pitched as a hub for civic engagement it has attracted Alt-right users since January 2021.

Theme 2: Technological

Table A:2 Open-source data for modelling

Database	Description
HatebaseTwitter	Combination of a hate speech lexicon taken from Hatebase and tweet history of 33,000 Twitter users (c.85m tweets). Via crowdsourcing, they annotated each tweet as hate speech, offensive (but not hate speech), or neither hate speech nor offensive. A commonly-used subset of this dataset is also available, containing 14,510 tweets. Hatebase is a collaborative, regionalised repository for multilingual hate speech that has amassed c.4k terms, in 98 languages across 178 countries.
WaseemA	A study that gathered c.17k tweets and labelled them as racist, sexist or neither, was developed from a long-list of 136k tweets.
WaseemB	The second set of 136k tweets was labelled by feminist and anti-racism activists.
Stormfront	A dataset developed from white supremacist site Stormfront annotated at sentence level and resulting in c.11k sentences labelled hate, no hate, relation or skip.
Trolling, Aggression and Cyberbullying (TRAC)	Combined Hindi/ English dataset that is publicly available and containing c. 16k Facebook comments labelled as overtly aggressive, covertly aggressive, or non-aggressive. The project also contains a smaller Twitter list of c.1,300 tweets.
HatEval	A multilingual set targeting hate speech in women and immigrants labelled on whether the tweet expressed hate towards women, whether it was aggressive, and whether the tweet was directed towards an individual or the entire group.
Kaggle[231]	A dataset consisting of c.9k social media posts that are labelled as insulting or not insulting.

[231] *Toxic Comment Classification Challenge* (no date) *Kaggle.com*. Available at: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge (Accessed: September 17, 2021).

	Kaggle is used by Conversation AI, a research initiative founded by Jigsaw and Google.
GermanTwitter	A c.540 tweet list focused on the German attitude to the refugee crisis.
Founta et al.[232] (2018)	Using Twitter and crowdsourcing techniques Founta et al. labelled c.80,000 abusive tweets. Labels included in the study include normal, spam, abusive and hateful.
HatEval[233] (2019)	Using Twitter and focusing on abuse against women and migrants, this is a multilingual model that works for both English and Spanish tweets.
Gao and Huang (2017)	A study that uses Fox News to determine hateful language in a context-aware model.
Fersini, Nozza and Rosso (2018)	A model that uses Twitter to identify misogyny.
Warner and Hirschberg (2012)	A model that uses Yahoo! (news articles) and the American Jewish Congress (offensive websites) as a source-identifying anti-Semitic, anti-immigrant, and racist sentiment.
Zhang et al. (2018)	Zhang et al. uses Twitter to identify hate speech using a convolution-GRU based deep neural network
Kumar et al. (2018)	Using Facebook and Twitter as a source Kumar et al. flagged both overt and covert aggression across platforms in both English and Hindi.
Wulczyn et al. (2017)	Wulcyzn et al. looked at Wikipedia and developed 100k human-labelled comments and 63m machine-labelled comments, determining that personal attacks on the platform are not the result of anonymous users, or because of the contribution from anonymous users.

[232] Founta, A.-M. *et al.* (2018) *Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior*, *Aaai.org*. Available at: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17909/17041 (Accessed: September 17, 2021).

[233] Basile, V. *et al.* (2019) "SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter," in *Proceedings of the 13th International Workshop on Semantic Evaluation*. Stroudsburg, PA, USA: Association for Computational Linguistics.

OLID (OffensEval)	Zampieri et al.[234] aimed to predict the type and target of offensive social media posts online using both Twitter and Facebook as a source, categorising tweets at three levels. ● Level A: offensive language detection (not offensive/ offensive) ● Level B: categorisation of offensive language (targeted insult/ untargeted) ● Level C: offensive language target (individual/ group/ other)
AbuseEval[235]	Using Twitter to identify both explicitly and implicitly offensive abuse, building on the OLID identification model. The outcome is a newly created resource, AbuseEval v1.0, which aims to address some of the existing issues in the annotation of offensive and abusive language (e.g., the explicitness of the message, presence of a target, need for context, and interaction across different phenomena).
SFU Opinion and Comments Corpus (SOCC)[236]	Using the Globe and Mail (online news sites) as a source Kolhatker et al. (2019) identified very toxic, toxic and mildly toxic comments over a five-year period, analysing c.10k articles and c.663k comments.
Razavi et al. (2010)	A weighted model that identifies offensive and abusive comments online.
Golbeck et a. (2017)[237]	Using Twitter to identify harassment messages. Offensive messages were identified manually across 35k tweets.

[234] Zampieri, M. *et al.* (2019) "Predicting the type and target of offensive posts in social media," in *Proceedings of the 2019 Conference of the North*. Stroudsburg, PA, USA: Association for Computational Linguistics.

[235] Caselli, T. *et al.* (2020) "I feel offended, don't be abusive! Implicit/explicit messages in offensive and abusive language." Available at: https://www.semanticscholar.org/paper/2b428c95b40d90f268dda2e12734ad5c154288b8 (Accessed: September 17, 2021).

[236] Kolhatkar, V. *et al.* (2020) "The SFU opinion and Comments Corpus: A corpus for the analysis of online news comments," *Corpus pragmatics*, 4(2), pp. 155–190.

[237] Golbeck, J. *et al.* (2017) "A large labeled corpus for online harassment research," in *Proceedings of the 2017 ACM on Web Science Conference*. New York, NY, USA: ACM.

Table A:3 Shared Task Organisation Checklist

Transparency	• Can organisers participate? • Can annotators participate • Can evaluators participate? • Is there a global mailing list to keep participants in the loop with responses to any questions or changes? • Will timelines be provided from the announcement of the shared task, and will they be publicly available? • Are timelines realistic and will unforeseen changes be communicated promptly? • What is the type of shared task (open/closed/both)? • Who is the invited participant audience (Research institution teams only? Industry teams also? Will there be separate tracks?) • Have the licensing conditions for sharing data been verified? • Where relevant, have steps been taken to safeguard privacy rights or sensitive data (e.g., anonymisation)?	• If manually annotated by crowd workers, were annotators fairly compensated? • Are there clear annotation guidelines for new datasets being used in the shared task and are they publicly available? • Is there consistency in annotation and format across multiple datasets? • Are details of any changes in data formats or annotation compared to previous tasks clearly communicated? • If an annotation tool is provided, are there one or two organisers representing technical support points of contact, should any issues be encountered? • Is there a (documented and accessible) quality control step before releasing datasets for the task? • How will systems be evaluated (remote VMs, web-based interfaces, APIs, etc.)?

Reporting and Replicability	• Is the requirement of a system description paper from each team made clear? • Is there a template for the system description to encourage consistency across participants' reporting? • Can shorter description papers from those ranking below a certain threshold be submitted? • Can longer description papers for complex systems be submitted? • Can system description papers be accompanied by supplementary materials? • Are participants encouraged to also provide an error analysis alongside negative results? • Is information clearly provided on whether or not an overview paper will be produced and how much detail it will contain?	• Is it clear to participants who should be named as authors on each publication (participants, annotators, etc.)? • Are participants required to release their code (including a detailed README) after the shared task? • Are participants required to release their data after the shared task (where copyright/licensing allows)? • Are participants required to release a detailed list of all data used in their system submissions? • Are other specific details made clear on the shared task website (e.g., participants that do not submit a description paper will not be ranked; participants using additional proprietary data will not be ranked, etc.)?
System Ranking	• Are participants encouraged to share negative results and can they choose not to be ranked if so preferred? • Can participants who rank low (below a chosen threshold) choose to be anonymised?	• Is it possible to report only the top N-ranked system scores in order to avoid withdrawal or fear of negative results?
Metrics	• Are evaluation metrics clearly stated when the shared task is announced? • Are all metrics to be used for ranking systems clearly stated?	• Will evaluation scripts will be shared after the shared task (or, if applicable, are they already available)?

Source: Escartín et al. (2021)

Theme 4: Legal, Political and Ethical

All-party Parliamentary Group's Trust, Transparency and Tech (2019)

- Recommendation 1: To build public confidence and acceptability, providers of public services should address ethics as part of their 'licence to operate'. A core principle should be that the public's views on data exploitation are proactively built into an ethical assessment at the service design stage.
- Recommendation 2: The citizen should be given access to simple and meaningful information, akin to the transparency principles underpinning Freedom of Information. This duty should apply to all those using data exploitation to deliver public services, as part of their 'licence to provide public services.
- Recommendation 3: The citizen should have a 'right to explanation', via a duty on all those delivering public services to provide easy to understand information on the factors taken into account in algorithm-based 'black-box' decisions as they affect the individual.
- Recommendation 4: There should be clear lines of accountability on data and algorithm use to the top of every organisation providing public services, including accessible complaints and redress processes. This could be achieved by extending the Data Protection Officer role and updating company director responsibilities.
- Recommendation 5: To ensure a consistent experience for the citizen, all Departments' existing governance arrangements should be assessed to ensure they are providing a coherent ethics framework for devolved public service delivery enforceable through respective regulators. Where necessary independent Data Ethics Advisory Boards should be established.
- Recommendation 6: An organisation should address the trust risks that could inhibit innovation Develop a user-friendly means - such as a kitemark - to show when a decision is taken by machine intelligence, and when you are interacting with a machine, not a human, and mandate its use across government and public service delivery in higher risk areas and provide central guidance on 'responsible trials' of Artificial Intelligence technologies such as biometrics and facial recognition as well as autonomous vehicles.
- Recommendation 7: Prioritise work on 'consent' as this is an aspect particularly challenged by data-driven technology, and carry out a full thematic review into a model for assumed public consent for common good, taking account of lessons learned. This should consider issues around informed versus implied consent, and how to ensure the consent process is fit for purpose and not a simple tick-box exercise.

Table A:4 Transdisciplinary best practice

Rule	Explanation
Develop reflexive habits	Exploration of one's own discipline through the lens of another.
Plan early and often	This means: ● Defining the purpose of the collaboration; ● Assigning roles and responsibilities for all collaboratory members involved in the project lifecycle, including principal investigators and team leads; ● Outlining benchmarks of success (i.e., project milestones); and ● Defining collaboration tools and how they relate to the purpose of the project, such as communication platforms and meeting schedules.
Speak the same language	Adopt an inclusive environment that encourages questions, fosters understanding of new concepts, and aids in vocabulary building.
Design the project so everyone benefits	Project planning should incorporate the research agendas of all collaborators. This should include: ● Create a flexible-by-design framework that can accommodate variable scope and unanticipated results. In other words, give room to both data scientists and disciplinary researchers to pursue what matters to them, while collaborating on the project; ● Specify the distinct contribution that each collaborator has to offer to their field; ● Identify inclusive objectives and/or outputs that allow each contributor to advance their own professional goals and research agendas; ● Account for differences in the fundamental approach to research between disciplines and practices, including methodology, experimental design, and analysis; ● Clarify that the results of collaborative research, including data science methods, will ultimately be evaluated by disciplinary experts; ● Do not assume that disciplinary contributions will contribute to the research portfolio of the data scientists,

	and vice versa; and ● Revise and improve your plan as you onboard new collaborators
Fail early and often	Data science projects involve perceived failures in the short term. Collaborative teams should fully embrace failure and learn to leverage these setbacks into opportunities for growth and success of their collaboration.
Share collaboration tools	Transdisciplinary collaboration should leverage the tools, skills, and resources that each member brings to the project by sharing them freely.
Manage your data like the collaboration depends on it	Collaborative teams should become proficient in data management best practices and work together to create data management plans that support FAIR data principles (Findable, Accessible, Interoperable, and Reusable).
Write code that can be re-used and reproduced	Strong code-writing skills foster ethical and responsible research outcomes, such as reproducibility
Observe ethical hygiene	Researchers should stay current on best practices, observe ethical hygiene throughout the research lifecycle, and prioritise the ethical guidelines published by their research sponsors.
Document collaboration	The experiences, reflections, and evolving best practices that result from data science collaborations can benefit the entire research community by providing anecdotal evidence about what works. Transdisciplinary teams should regularly document their collaborative experiences, regardless of perceived successes or failures.

Source: Sahneh et al. (2021)

Citations

- Aiken, M. P. (2016). The Cyber Effect. New York. Random House, Spiegel & Grau.
- Aiken, M., 2021. Manipulating Fast, and Slow. [online
- Aiken, M., 2021. Mass Killing and Technology: The Hidden Links. [online
- Ali, S., Saeed, M.H., Aldreabi, E., Blackburn, J., De Cristofaro, E., Zannettou, S. and Stringhini, G. (2021). Understanding the Effect of Deplatforming on Social Networks. 13th ACM Web Science Conference 2021. [online
- Allison, P. R. (2019) Politics, privacy and porn: the challenges of age-verification technology, Computerweekly.com. ComputerWeekly.com. Available at: https://www.computerweekly.com/feature/Politics-privacy-and-porn-the-challenges-of-age-verification-technology (Accessed: September 17, 2021).
- Al-Mansoori, R. S. et al. (2021) Digital Wellbeing for All: Expanding Inclusivity to Embrace Diversity in Socio-Emotional Status, Researchgate.net. Available at: https://www.researchgate.net/profile/Raian-Ali/publication/353526705_Digital_Wellbeing_for_All_Expanding_Inclusivity_to_Embrace_Diversity_in_Socio-Emotional_Status/links/6101c0461e95fe241a95ba2e/Digital-Wellbeing-for-All-Expanding-Inclusivity-to-Embrace-Diversity-in-Socio-Emotional-Status.pdf (Accessed: September 17, 2021).
- Alvernia Online. (2021). Group Polarization in Social Psychology | Alvernia Online. [online
- Anon, (2018). General Aggression Model. [online
- Anti-Defamation League (2020) Free to Play? Hate, Harassment and Positive Social Experiences in Online Games 2020, Adl.org. Available at: https://www.adl.org/media/15349/download (Accessed: September 17, 2021).
- Anti-Defamation League. (n.d.). Disruption and Harms in Online Gaming Framework. [online
- Atske, S. (2021) The state of online harassment, Pewresearch.org. Available at: https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/ (Accessed: September 17, 2021).
- Atske, S. (2021a) Social media use in 2021, Pewresearch.org. Available at: https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/ (Accessed: September 17, 2021).
- Australia Government (no date) Doxing trends and challenges — position statement Gov.au. Available at: https://www.esafety.gov.au/about-us/tech-trends-and-challenges/doxing (Accessed: September 17, 2021).
- Awan, I. and Zempi, I. (2016) “The affinity between online and offline anti-Muslim hate crime: Dynamics and impacts,” Aggression and violent behavior, 27, pp. 1–8.
- Awan, I., Sutch, H. and Carter, P. (2019). Extremism Online - Analysis of extremist material on social media. [online
- Barrett, P. (2020). Who Moderates the Social Media Giants? [online
- Barrett, P.M., Hendrix, J. and Grant Sims, J. (2021). Polarization Report. [online
- Basile, V. et al. (2019) “SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter,” in Proceedings of the 13th International Workshop on Semantic Evaluation. Stroudsburg, PA, USA: Association for Computational Linguistics.
- Beauchamp, Nick, Ioana Panaitiu and Spencer Piston (2018) “Trajectories of Hate: Mapping Individual Racism and Misogyny on Twitter.” Unpublished Working Paper.
- Benigni, M. C., Joseph, K. and Carley, K. M. (2017) “Online extremism and the communities that sustain it: Detecting the ISIS supporting community on Twitter,” PloS one, 12(12), p. e0181405.
- Beres, N. A. et al. (2021) “Don’t you know that you’re toxic: Normalization of toxicity in online gaming,” in Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM.
- Bernal, N. (2021) Facebook’s content moderators are fighting back, WIRED UK. Available at: https://www.wired.co.uk/article/facebook-content-moderators-ireland (Accessed: September 17, 2021).
- Bernal, N. (2021) Facebook’s content moderators are fighting back, WIRED UK. Available at: https://www.wired.co.uk/article/facebook-content-moderators-ireland (Accessed: September 17, 2021).
- Blackman, R. (2020). A Practical Guide to Building Ethical AI. Harvard Business Review. [online
- Bond, S. (2021) “Fast-Growing Alternative to Facebook and Twitter Finds Post-Trump Surge ‘Messy,’” NPR. Available at: https://www.npr.org/2021/01/22/958877682/fast-growing-alternative-to-facebook-twitter-finds-right-wing-surge-messy?t=1630676748964 (Accessed: September 17, 2021).
- Breland, A. (2020). Facebook announces crackdown on QAnon, antifa, and militias. [online
- Buerger, C. and Wright, L. (2019) Counterspeech: A Literature Review, Dangerousspeech.org. Available at: https://dangerousspeech.org/wp-content/uploads/2019/11/Counterspeech-lit-review_complete-11.20.19-2.pdf (Accessed: September 17, 2021).

- Buntz, B. (2020) 2020 predictions: Computer vision projects will gain ground, Iotworldtoday.com. Available at: https://www.iotworldtoday.com/2020/01/06/2020-predictions-computer-vision-projects-will-gain-ground/ (Accessed: September 17, 2021).
- Buntz, B. (2020) 2020 predictions: Computer vision projects will gain ground, Iotworldtoday.com. Available at: https://www.iotworldtoday.com/2020/01/06/2020-predictions-computer-vision-projects-will-gain-ground/ (Accessed: September 17, 2021).
- Calvo, R. and Peters, D. (n.d.). About. [online
- Canales, K. (2021). Facebook's AI moderation reportedly can't interpret many languages, leaving users in some countries more susceptible to harmful posts. [online
- Carlsson, K. (2019). The Forrester New Wave™: Computer Vision Platforms, Q4 2019 The 11 Providers That Matter Most and How They Stack Up [online
- Casciani, D. and De Simone, D. (2021). Incels: A new terror threat to the UK? BBC News. [online
- Caselli, T. et al. (2020) "I feel offended, don't be abusive! Implicit/explicit messages in offensive and abusive language." Available at: https://www.semanticscholar.org/paper/2b428c95b40d90f268dda2e12734ad5c154288b8 (Accessed: September 17, 2021).
- Castaño-Pulgarín, S. A. et al. (2021) "Internet, social media and online hate speech. Systematic review," Aggression and violent behavior, 58(101608), p. 101608.
- Chandrasekharan, E. et al. (2017) "You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech," Proceedings of the ACM on human-computer interaction, 1(CSCW), pp. 1–22.
- Chen, M., Cheung, A. S. Y. and Chan, K. L. (2019) "Doxing: What adolescents look for and their intentions," International journal of environmental research and public health, 16(2). doi: 10.3390/ijerph16020218.
- Chetty, N. and Alathur, S. (2018) "Hate speech review in the context of online social networks," Aggression and violent behavior, 40, pp. 108–118.
- Chi, L. and Zhu, X. (2017) "Hashing techniques: A survey and taxonomy," ACM computing surveys, 50(1), pp. 1–36.
- Cho, D. and Kwon, H. (2015) The impacts of identity verification and disclosure of social cues on flaming in online user comments, Researchgate.net.
- Cinelli, M., de Francisci Morales, G., Galeazzi, A., Quattrociocchi, W. and Starnini, M., 2021. The echo chamber effect on social media. [online
- Classroom.cloud.(no date), eSafety/Safeguarding – A helping hand, Available at: https://classroom.cloud/online-safety/ (Accessed: September 17, 2021).
- Clegg, N. (2020). Facebook Does Not Benefit from Hate. [online
- Coldewey, D. (2019) "Snopes rolls its own crowdfunding infrastructure to prepare for 2020's disinformation warfare," TechCrunch, 20 December. Available at: http://techcrunch.com/2019/12/20/snopes-rolls-its-own-crowdfunding-infrastructure-to-prepare-for-2020s-disinformation-warfare/ (Accessed: September 17, 2021).
- Cory, B. Y. N. (no date) How website blocking is curbing digital piracy without "breaking the internet," Gov.pt. Available at: https://www.igac.gov.pt/documents/20178/557437/Estudo_2017/3adcf3b7-e9ca-497a-bebd-5fc72cec72e7 (Accessed: September 17, 2021).
- Costa, E. and Halpern, D. (2019) The behavioural science of online harm and manipulation, and what to do about it, Cxmlab.com. Available at: https://www.cxmlab.com/wp-content/uploads/2019/07/BIT_The-behavioural-science-of-online-harm-and-manipulation-and-what-to-do-about-it_Single-2.pdf (Accessed: September 17, 2021).
- Costa, E. and Halpern, D. (2019) The behavioural science of online harm and manipulation, and what to do about it, Cxmlab.com. Available at: https://www.cxmlab.com/wp-content/uploads/2019/07/BIT_The-behavioural-science-of-online-harm-and-manipulation-and-what-to-do-about-it_Single-2.pdf (Accessed: September 17, 2021).
- Costello, Matthew and Hawdon.(2018) "Who Are the Online Extremists Among Us? Sociodemographic Characteristics, Social Networking, and Online Experiences of Those Who Produce Online Hate Materials." Violence and Gender 5(1):55–60.
- Crockett, M. (2016) The internet (never) forgets, Wpmucdn.com. Available at: https://cpb-us-w2.wpmucdn.com/smulawjournals.org/dist/8/7/files/2018/11/4_The-Internet-Never-Forgets.pdf (Accessed: September 17, 2021).
- Curtis, A. (2020). About the Safety Tech Innovation Network. SafetyTech Innovation Network. [online
- Danit, G. I. G., Alves, T. and Martinez, G. (2015) Countering online hate speech, Unesco.org. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000233231. (Accessed: September 17, 2021).
- de la Garza, A. (2021). What Would a Climate-Conscious Facebook Look Like? [online

- DE SATGE, F. (2021). The Central Role of Memes on Alt-Right Radicalisation in the “Chanosphere.” [online
- Department for Digital, Culture, Media & Sport (2019). Online Harms White Paper. [online
- Department for Digital, Culture, Media & Sport (2021) Understanding and reporting online harms on your online platform, Gov.uk. Available at: https://www.gov.uk/guidance/understanding-and-reporting-online-harms-on-your-online-platform (Accessed: September 17, 2021).
- Department for Digital, Culture, Media & Sport. (2020). Government publishes new strategy to kickstart data revolution across the UK. [online
- Department for Digital, Culture, Media & Sports (2020) Safer technology, safer users: The UK as a world-leader in Safety Tech, Gov.uk. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/974414/Safer_technology__safer_users-_The_UK_as_a_world-leader_in_Safety_Tech_V2.pdf (Accessed: September 17, 2021).
- Department for Digital, Culture, Media & Sports (2020) Safer technology, safer users: The UK as a world-leader in Safety Tech, Gov.uk. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/974414/Safer_technology__safer_users-_The_UK_as_a_world-leader_in_Safety_Tech_V2.pdf (Accessed: September 17, 2021).
- DeStreel, A., Defreyne, E., Jacquemin, H., Ledger, M. and Michel, A. (2020). Online Platforms’ Moderation of Illegal Content Online Law, Practices and Options for Reform. [online
- Deutsche Welle (www. dw.com) (2020) US: Trump fans choose Parler over Twitter, Www.dw.com. Deutsche Welle (www.dw.com). Available at: https://www.dw.com/en/donald-trump-twitter-parler-free-speech/a-55582802 (Accessed: September 17, 2021).
- Dilmegani, C. (2021). The Ultimate Guide to Synthetic Data in 2020. [online
- disinfocloud.com. (n.d.). Disinfo Cloud. [online
- Disinformation Index. (2020). Bankrolling Bigotry: An Overview of the Online Funding Strategies of American Hate Groups. [online
- Douglas, M. (2020). Media companies can now be held responsible for your dodgy comments on social media. [online
- Eckert, S. and Metzger-Riftkin, J. (2020) “Doxxing,” The International Encyclopedia of Gender, Media, and Communication. Wiley, pp. 1–5. doi: 10.1002/9781119429128.iegmc009.
- Ehrenkranz, M. (2018) Facebook is using new AI tools to detect child porn and catch predators, Gizmodo. Available at: https://gizmodo.com/facebook-is-using-new-ai-tools-to-detect-child-porn-and-1829968486 (Accessed: September 17, 2021).
- Electronic Arts (2020) The Positive Play Charter, Www.ea.com. Available at: https://www.ea.com/en-gb/news/the-positive-play-charter (Accessed: September 17, 2021).
- European Commission (2021) Terrorist content online, Europa.eu. Available at: https://ec.europa.eu/home-affairs/system/files/2021-05/202104_terrorist-content-online_en.pdf (Accessed: September 17, 2021).
- European Commission. (2021). The Digital Services Act package | Shaping Europe’s digital future. [online
- European Digital Rights (EDRi). (2020). French Avia law declared unconstitutional: what does this teach us at EU level? [online
- European Internet Forum. (2021). European Internet Forum - Decrypting the encryption debate: How to ensure public safety with a privacy-preserving and secure Internet? [online
- Facebook, (no date) Use Keyword Alerts to spot when specific terms are used in your group, Facebook.com. Available at: https://www.facebook.com/community/whats-new/using-keyword-alerts/ (Accessed: September 17, 2021).
- Farid, H. and Schindler, H.-J. (2020). Deep Fakes on the Threat of Deep Fakes to Democracy and Society. [online
- Felmlee, D. et al. (2020) “Can social media anti-abuse policies work? A quasi-experimental study of online sexist and racist slurs,” Socius : sociological research for a dynamic world, 6, p. 237802312094871.
- Felmlee, D. et al. (2020) “Can social media anti-abuse policies work? A quasi-experimental study of online sexist and racist slurs,” Socius : sociological research for a dynamic world, 6, p. 237802312094871.
- Felzmann, H., Fosch-Villaronga, E., Lutz, C. and Tamò-Larrieux, A. (2020). Towards Transparency by Design for Artificial Intelligence. Science and Engineering Ethics, 26(6), pp.3333–3361.
- Fiss, J.M., Joelle (2019). Germany’s Online Crackdowns Inspire the World’s Dictators. [online
- Founta, A.-M. et al. (2018) Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior, Aaai.org. Available at:

https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17909/17041 (Accessed: September 17, 2021).
- Fraga-Lamas, P. and Fernández-Caramés, T. M. (2019) “Fake news, disinformation, and deepfakes: Leveraging Distributed Ledger Technologies and blockchain to combat digital deception and counterfeit reality,” arXiv [cs.CY
- Francis, G. (2020). Safety tech at the UN. [online
- Francis, G. (2021). New industry consortium launches to transform access to online harms data. [online
- Gagliardone, I. (2015) Countering Online Hate Speech - UNESCO. UNESCO Publishing.
- Germani, F. and Biller-Andorno, N. (2021). The anti-vaccination infodemic on social media: A behavioral analysis. PLOS ONE, [online
- Gibbons, V.-M. (2020). Celebrating Startup Success at Facebook Accelerator London. [online
- Giles, K. and Mustaffa, M. (2019). The Role of Deepfakes in Malign Influence Campaigns. [online
- Golbeck, J. et al. (2017) “A large labeled corpus for online harassment research,” in Proceedings of the 2017 ACM on Web Science Conference. New York, NY, USA: ACM.
- Gorwa, R., Binns, R. and Katzenbach, C. (2020) “Algorithmic content moderation: Technical and political challenges in the automation of platform governance,” Big data & society, 7(1), p. 205395171989794.
- Gorwa, R., Binns, R. and Katzenbach, C. (2020) “Algorithmic content moderation: Technical and political challenges in the automation of platform governance,” Big data & society, 7(1), p. 205395171989794.
- Gorwa, R., Binns, R. and Katzenbach, C. (2020) “Algorithmic content moderation: Technical and political challenges in the automation of platform governance,” Big data & society, 7(1), p. 205395171989794.
- Government Office for Science. (2020). The future of citizen data systems. [online
- Graves, L. (2018) Understanding the promise and limits of automated fact-checking, Ox.ac.uk. Available at: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2018-02/graves_factsheet_180226%20FINAL.pdf (Accessed: September 17, 2021).
- Greaves, M. (2020). “Deepfakes” ranked as most serious AI crime threat. [online
- Guterres, A. (2019b). United Nations Strategy and Plan of Action on Hate Speech. [online
- Harel, T.O., Jameson, J.K. and Maoz, I. (2020). The Normalization of Hatred: Identity, Affective Polarization, and Dehumanization on Facebook in the Context of Intractable Political Conflict. Social Media + Society, 6(2), p.205630512091398.
- Harel, T.O., Jameson, J.K. and Maoz, I. (2020b). The Normalization of Hatred: Identity, Affective Polarization, and Dehumanization on Facebook in the Context of Intractable Political Conflict. Social Media + Society, 6(2), p.205630512091398.
- Harwell, D. and Dou, E. (2020). Huawei tested AI software that could recognize Uighur minorities and alert police, report says. Washington Post. [online
- Haselton, T. and Graham, M. (2019) About 2,200 people watched the German synagogue shooting on Amazon’s Twitch, CNBC. Available at: https://www.cnbc.com/2019/10/09/the-german-synagogue-shooting-was-streamed-on-twitch.html (Accessed: September 17, 2021).
- Hatelab. (n.d.). HateLab – A global repository for data and insight into hate crime and speech. [online
- Hawdon, James, Atte Oksanen and Pekka Ras¨anen. 2014. “Victims of online hate groups: American youths exposure to online hate speech.” The causes and consequences of group violence: From bullies to terrorists pp. 165–182.
- HEIKKILÄ, M. (2021). European Parliament calls for a ban on facial recognition. [online
- Heller, B. (2019) Combating Terrorist-Related Content Through AI and Information Sharing, Annenbergpublicpolicycenter.org. Available at: https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/05/Combating_Terrorist_Content_TWG_Heller_April_2019.pdf (Accessed: September 17, 2021).
- Internet Watch Foundation (2019) Annual Report 2019, iwf.org.uk/. Available at: https://www.iwf.org.uk/sites/default/files/reports/2020-04/IWF_Annual_Report_2020_Low-res-Digital_AW_6mb.pdf (Accessed: September 17, 2021).
- Internet Watch Foundation (2020) Face the facts The Annual Report 2020, www.iwf.org.uk/. Available at: https://www.iwf.org.uk/sites/default/files/inline-files/PDF%20of%20IWF%20Annual%20Report%202020%20FINAL%20reduced%20file%20size.pdf (Accessed: September 17, 2021).
- Internet Watch Foundation (no date) Domain Alerts, Org.uk. Available at: https://www.iwf.org.uk/our-services/domain-alerts (Accessed: September 17, 2021).
- Internetsociety.org.(no date) An overview of Internet content blocking Available at: https://www.internetsociety.org/resources/doc/2017/internet-content-blocking/ (Accessed: September 17, 2021).

- Issie Lapowsky (2021). OnlyFans reveals Visa and MasterCard's hold on online speech. [online
- IWF. (2020). "Game-changing" chatbot to target people trying to access child sexual abuse online. [online
- Izsak, R. (2015) "Hate speech and incitement to hatred against minorities in the media." UN Humans Rights Council.
- Jhaver, S., Boylston, C., Yang, D. and Bruckman, A. (2021). Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), pp.1–30.
- Kahn, J. (2021). Be afraid: Executives warn about personal data harvesting and use. [online
- Kathryn Tremlett (2021) A sit down with report harmful content, Org.uk. Available at: https://swgfl.org.uk/magazine/a-sit-down-with-report-harmful-content/ (Accessed: September 17, 2021).
- Keller, F. B. et al. (2020) "Political astroturfing on twitter: How to coordinate a disinformation campaign," Political communication, 37(2), pp. 256–280.
- Kelly, M., DiBranco, A. and DeCook, J.R. (2021). Mass Violence and Terrorism since Santa Barbara. New America. Available at:https://d1y8sb8igg2f8e.cloudfront.net/documents/Misogynist_Incels_and_Male_Supremacism.pdf.
- Kolhatkar, V. et al. (2020) "The SFU opinion and Comments Corpus: A corpus for the analysis of online news comments," Corpus pragmatics, 4(2), pp. 155–190.
- Kor-Sins, R. (2021). The alt-right digital migration: A heterogeneous engineering approach to social media platform branding. New Media & Society, p.146144482110388.
- L1ght (2020). Rising Levels of Hate Speech & Online Toxicity During This Time of Crisis. [online
- Land, M. K. and Hamilton, R. J. (2020) "Beyond takedown: Expanding the toolkit for responding to online hate," SSRN Electronic Journal. doi: 10.2139/ssrn.3514234.
- LaRose, R. (2015) "The psychology of interactive media habits," in The Handbook of the Psychology of Communication Technology. Chichester, UK: John Wiley & Sons, Ltd, pp. 365–383.
- Lewis, J. and Marsden, S. (2021). Countering Violent Extremism Interventions: Contemporary Research. [online
- Licklider. (2009). Man-Computer Symbiosis. Available at: http://worrydream.com/refs/Licklider%20-%20Man-Computer%20Symbiosis.pdf [Accessed September 17, 2021
- Logan, S. (2018). Facebook and Vietnam's new cybersecurity law. [online
- Lyngs, U. et al. (2020) "I just want to hack myself to not get distracted: Evaluating design interventions for self-control on Facebook," in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM.
- MacAvaney, S. et al. (2019) "Hate speech detection: Challenges and solutions," PloS one, 14(8), p. e0221152.
- Mamié, R., Horta Ribeiro, M. and West, R. (2021). Are Anti-Feminist Communities Gateways to the Far Right? Evidence from Reddit and YouTube. 13th ACM Web Science Conference 2021.
- Marantz, A. (2019). Reddit and the Struggle to Detoxify the Internet. [online
- Margetts, H., Vidgen, B. and Burden, E. (2021) VSP Regulation and the broader context, Org.uk. Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0022/216490/alan-turing-institute-report-understanding-online-hate.pdf (Accessed: September 17, 2021).
- Maschmeyer, L., Deibert, R.J. and Lindsay, J.R. (2020b). A tale of two cybers - how threat reporting by cybersecurity firms systematically underrepresents threats to civil society. Journal of Information Technology & Politics, 18(1), pp.1–20.
- Mathew, B. et al. (2018) "Spread of hate speech in online social media," arXiv [cs.SI
- Mathew, B., Dutt, R., Goyal, P. and Mukherjee, A. (2019). Spread of Hate Speech in Online Social Media. [online
- Matjašič, P. (2021). Spanish gag law: The original sin and ongoing penance. [online
- McCauley, C. and Moskalenko, S. (2008) "Mechanisms of political radicalization: Pathways toward terrorism," Terrorism and political violence, 20(3), pp. 415–433.
- McGuire, K. (2020) The Shady Side of Twitch, Looper.com. Looper. Available at: https://www.looper.com/263700/the-shady-side-of-twitch/ (Accessed: September 17, 2021).
- Mehrey, A. and Bharath (2021). Facebook Startup Funding | Startups Funded by the Facebook. [online
- mint. (2021). Facebook says it has spent $13 bn on safety, security since 2016 US election. [online
- Montgomery, M. (2020) Disinformation as a wicked problem: Why we need co-regulatory frameworks, Brookings.edu. Available at: https://www.brookings.edu/wp-content/uploads/2020/08/Montgomery_Disinformation-Regulation_PDF.pdf (Accessed: September 17, 2021).
- Moonshot (no date) The Redirect Method: How it works (no date) Moonshotteam.com. Available at: https://moonshotteam.com/redirect-method/ (Accessed: September 17, 2021).

- Mossie, Z. and Wang, J.-H. (2020) “Vulnerable community identification using hate speech detection on social media,” Information processing & management, 57(3), p. 102087.
- Munn, L. (2020) “Angry by design: toxic communication and technical architectures,” Humanities and Social Sciences Communications, 7(1), pp. 1–11.
- Murphy, H. and Yang, Y. (2019) “TikTok rushes to build moderation teams as concerns rise over content,” Irish times, 20 December. Available at: https://www.irishtimes.com/business/technology/tiktok-rushes-to-build-moderation-teams-as-concerns-rise-over-content-1.4121460 (Accessed: September 17, 2021).
- Murphy, H. and Yang, Y. (2019) “TikTok rushes to build moderation teams as concerns rise over content,” Irish times, 20 December. Available at: https://www.irishtimes.com/business/technology/tiktok-rushes-to-build-moderation-teams-as-concerns-rise-over-content-1.4121460 (Accessed: September 17, 2021).
- Nast, C. (2021). Facebook’s content moderators are fighting back. [online
- Newton, C. (2019) Facebook moderators break their NDAs to expose desperate working conditions, The Verge. Available at: https://www.theverge.com/2019/6/19/18681845/facebook-moderator-interviews-video-trauma-ptsd-cognizant-tampa (Accessed: September 17, 2021).
- Nilson, G. (2021). Litecoin and Walmart. [online
- OFCOM (2019) Use of AI in Online Content Moderation www.ofcom.org.uk. Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf (Accessed: September 17, 2021).
- Ofcom. (2019). Online market failures and harms: An economic perspective on the challenges and opportunities in regulating online services. [online
- Onix (2021) How TikTok has changed live streaming for social media. Onix-systems.com. Available at: https://onix-systems.com/blog/how-did-tiktok-social-media-live-streaming-change-everything (Accessed: September 17, 2021).
- Online Harms White Paper: Full government response to the consultation (2020) Gov.uk. Available at: https://www.gov.uk/government/consultations/online-harms-white-paper/outcome/online-harms-white-paper-full-government-response (Accessed: September 17, 2021).
- Osborne Clarke (2020) Online harms: The new legal framework for addressing “hate speech” in France and in Germany Osborneclarke.com. Available at: https://www.osborneclarke.com/insights/online-harms-new-legal-framework-addressing-hate-speech-france-germany/ (Accessed: September 17, 2021).
- OSTIA. (2021). OSTIA - Online Safety Tech Industry Association. [online
- Packer, B. (2021). Online Harms: A comparative analysis. [online
- Pandith, F. and Ware, J. (2021) Teen terrorism inspired by social media is on the rise. Here’s what we need to do, NBC News. Available at: https://www.nbcnews.com/think/opinion/teen-terrorism-inspired-social-media-rise-here-s-what-we-ncna1261307 (Accessed: September 17, 2021).
- Pandith, F., Ware, J. and Bloom, M. (2020) Female extremists in QAnon and ISIS are on the rise. We need a new strategy to combat them, NBC News. Available at: https://www.nbcnews.com/think/opinion/female-extremists-qanon-isis-are-rise-we-need-new-strategy-ncna1250619 (Accessed: September 17, 2021).
- Parliament (2020) Written evidence submitted by Twitter (COR0177) Parliament.uk. Available at: https://committees.parliament.uk/writtenevidence/5814/pdf/ (Accessed: September 17, 2021).
- Paul, K. (2019). Facebook’s crackdown on dangerous content in groups could backfire, experts say. [online
- Pelley, S., 2021. Whistleblower: Facebook is misleading the public on progress against hate speech, violence, misinformation. [online
- Perez, S. (2021). Facebook rolls out new tools for Group admins, including automated moderation aids. [online
- Peters, D. (2019). Beyond principles: A Process for Responsible Tech. [online
- Place, N. (2021) “Fake news got more engagement than real news on Facebook in 2020, study says,” Independent, 5 September. Available at: https://www.independent.co.uk/news/world/americas/fake-news-facebook-misinformation-study-b1914650.html (Accessed: September 17, 2021).
- Polger, D. R. (2018) Why we need more online friction, Techonomy.com. Available at: https://techonomy.com/2018/12/need-online-friction/ (Accessed: September 17, 2021).
- Policy Department for Citizens’ Rights and Constitutional Affairs (2020) The impact of algorithms for online content filtering or moderation, Europa.eu. Available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/657101/IPOL_STU(2020)657101_EN.pdf (Accessed: September 17, 2021).
- Priya, B. (2021). Private AI: Machine Learning on Encrypted Data. [online

- Ray, S. (2021) “The far-right is flocking to these alternate social media apps — not all of them are thrilled,” Forbes Magazine, 14 January. Available at: https://www.forbes.com/sites/siladityaray/2021/01/14/the-far-right-is-flocking-to-these-alternate-social-media-apps---not-all-of-them-are-thrilled/ (Accessed: September 17, 2021).
- Reddit apologises for online Boston “witch hunt.” (2013). BBC News. [online
- Reed, A. and Aryaeinejad, K. (2021). 2020 Trends in Terrorism: From ISIS Fragmentation to Lone-Actor Attacks. [online
- Rephrain. (2020). National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online. [online
- Reuters Institute for the Study of Journalism. (2021). Digital News Report 2021. [online
- Reuters (2021) “GoDaddy terminates hosting of Texas anti-abortion tip website,” 3 September. Available at: https://www.reuters.com/world/us/godaddy-terminate-hosting-texas-anti-abortion-tip-website-2021-09-03/ (Accessed: September 17, 2021).
- Richardson, Z. (2021). Online Safety and Media Regulation Bill: Social media firms facing hefty fines and criminal liability if they fail to meet new online safety standards. [online
- Richardson, Z. (2021). Online Safety and Media Regulation Bill: Social media firms facing hefty fines and criminal liability if they fail to meet new online safety standards. [online
- Ryan, C. D. et al. (2020) “Monetizing disinformation in the attention economy: The case of genetically modified organisms (GMOs),” European management journal, 38(1), pp. 7–18.
- Sabat, B. O., Ferrer, C. C. and Giro-i-Nieto, X. (2019) “Hate speech in pixels: Detection of offensive memes towards automatic moderation,” arXiv [cs.MM
- Safety Tech Innovation Network. (2021). German safety tech industry gains momentum. [online
- Sahneh, F., Balk, M.A., Kisley, M., Chan, C., Fox, M., Nord, B., Lyons, E., Swetnam, T., Huppenkothen, D., Sutherland, W., Walls, R.L., Quinn, D.P., Tarin, T., LeBauer, D., Ribes, D., Birnie, D.P., Lushbough, C., Carr, E., Nearing, G. and Fischer, J. (2021). Ten simple rules to cultivate transdisciplinary collaboration in data science. PLOS Computational Biology, 17(5), p.e1008879.
- Schuler, M. and Znaty, B. (2020). New law to fight online hate speech (Avia law) to reshape notice, take down and liability rules in France. [online
- Scott, J. and Savov, V. (2021). Australian Law Could Force Facebook, Google to Strip Content. Bloomberg.com. [online
- Scott, M. and Kayali, L. (2020) What happened when humans stopped managing social media content, POLITICO. Available at: https://www.politico.eu/article/facebook-content-moderation-automation/ (Accessed: September 17, 2021).
- Seering, J. (2020) “Reconsidering self-moderation: The role of research in supporting community-based models for online content moderation,” Proceedings of the ACM on human-computer interaction, 4(CSCW2), pp. 1–28.
- Sen, A. and Zadrozny, B. (2020). QAnon groups have millions of members on Facebook, documents show. [online
- Siegel, A. A. and Badaan, V. (2020) #No2Sectarianism: Experimental Approaches to Reducing Sectarian Hate Speech Online. (Accessed: September 17, 2021).
- Singh, S. (2019). Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content. [online
- Slane, A. (2007). ‘Democracy, social space and the Internet’, University of Toronto Law Journal, 57: 81 - 104
- Slattery, L. (2021). “Wild West” social media firms criticised for response to harmful content. [online
- Smith, A., (2021). Risk Factors and Indicators Associated with Radicalization to Terrorism in the United States: What Research Sponsored by the National Institute of Justice Tells Us. [online
- Snopes (2019) If Facebook is dealing with deceptive ‘BL’ network, it’s not working Snopes.com. Available at: https://www.snopes.com/news/2019/12/13/facebook-bl-cib/ (Accessed: September 17, 2021).
- Solove, D.J. (2006). A Taxonomy of Privacy. [online
- Soral, W., Liu, J. and Bilewicz, M. (2020) “Media of contempt: Social media consumption predicts normative acceptance of anti-Muslim hate speech and Islamoprejudice.” doi: 10.4119/IJCV-3774.
- Spiegel, J. (2018). Germany’s Network Enforcement Act and its impact on social networks. [online
- Spitaletta, J. A. and Hopkins, J. (2021) Operational Cyberpsychology: Adapting a Special Operations Model for Cyber Operations, Nsiteam.com. Available at: https://nsiteam.com/social/wp-content/uploads/2021/07/Invited-Perspective-Operational-Cyber-Psych_FINAL.pdf (Accessed: September 17, 2021).
- Staal, M. A. and Stephenson, J. A. (2013) “Operational psychology post-9/11: A decade of evolution,” Military psychology: the official journal of the Division of Military Psychology, American Psychological Association, 25(2), pp. 93–104.

- Statista. (2018). Facebook: annual revenue 2018 | Statistic. [online
- Stewart, E. (2020) America's growing fake news problem, in one chart, Vox. Available at: https://www.vox.com/policy-and-politics/2020/12/22/22195488/fake-news-social-media-2020 (Accessed: September 17, 2021).
- Study: On Twitter, false news travels faster than true stories (2018) Mit.edu. Available at: https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308 (Accessed: September 17, 2021).
- Tarnoff, B. (2016) "The Attention Merchants review – how the web is being debased for profit," The guardian, 26 December. Available at: http://www.theguardian.com/books/2016/dec/26/the-attention-merchants-tim-wu-review (Accessed: September 17, 2021).
- The Knights Foundation (2019). Disinformation, "fake news" and influence Campaigns on Twitter. [online
- The Law Commission. (2014). Hate Crime: Should the Current Offences be Extended? [online
- The New York Times. 2021. Facebook Dials Down the Politics for Users. [online
- Tindale, J. and Muirhead, O. (2019). Trust, Transparency and Technology: Building Data Policies for the Public Good. [online
- Torok, R. (2013) "Developing an explanatory model for the process of online radicalisation and terrorism," Security informatics, 2(1), p. 6.
- Townsend, M. (2021). How far right uses video games and tech to lure and radicalise teenage recruits. [online
- Toxic Comment Classification Challenge (no date) Kaggle.com. Available at: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge (Accessed: September 17, 2021).
- Trotzek, M., Koitka, S. and Friedrich, C. M. (2020) "Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences," IEEE transactions on knowledge and data engineering, 32(3), pp. 588–601.
- Trust & Safety Professional Association. (2021). Advancing the trust and safety profession through a shared community of practice. [online
- Twitter (2020) Twitter acquires Fabula AI to strengthen its machine learning expertise, Twitter.com. Available at: https://blog.twitter.com/en_us/topics/company/2019/Twitter-acquires-Fabula-AI (Accessed: September 17, 2021).
- Twitter (2020) Permanent suspension of @realDonaldTrump .Twitter.com. Available at: https://blog.twitter.com/en_us/topics/company/2020/suspension (Accessed: September 17, 2021).
- Ullmann, S. and Tomalin, M. (2020) "Quarantining online hate speech: technical and ethical perspectives," Ethics and information technology, 22(1), pp. 69–80.
- USA Today (2021) "Exclusive: 43% of Americans say a specific organization or people to blame for COVID-19." Available at: https://eu.usatoday.com/story/news/politics/2021/03/21/poll-1-4-americans-has-seen-asians-blamed-covid-19/4740043001/ (Accessed: September 17, 2021).
- Van Dorpe, S. (2021). Germany shows EU the way in curbing Big Tech. [online
- Van Royen, K. et al. (2017) "'Thinking before posting?' Reducing cyber harassment on social networking sites through a reflective message," Computers in human behavior, 66, pp. 345–352.
- Venkataramakrishnan, S. (2021). Online privacy: a fraught philosophical debate. [online
- Verfassungsblog. 2021. The UK's Online Safety Bill: Safe, Harmful, Unworkable?. [online
- Vidgen, B., Burden, E. and Margetts, H., (2021). Alan Turing Institute Understanding Online Hate VSP regulation and the broader context. [online
- Vidgen, B., Burden, E. and Margetts, H., (2021). Understanding Online Hate VSP regulation and the broader context. [online
- Vidgen, B., Burden, E. and Margetts, H., 2021. Understanding Online Hate VSP regulation and the broader context. [online
- VoCO (Verification of Children Online) Phase 2 Report (2020) Gov.uk. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/934131/November_VoCO_report_V4__pdf.pdf (Accessed: September 17, 2021).
- von Behr, I., Reding, A., Edwards, C. and Gribbon, L. (n.d.). Radicalisation in the digital era The use of the internet in 15 cases of terrorism and extremism. RAND Corporation.
- We Forum (2021) Big tech cannot crack down on online hate alone Weforum.org. Available at: https://www.weforum.org/agenda/2021/04/big-tech-cannot-crack-down-on-online-hate-alone/ (Accessed: September 17, 2021).
- Wijeratne, Y. (2020). Facebook, language and the difficulty of moderating hate speech | Media@LSE. [online
- Williams, M. (2019). Hatred Behind the Screens A Report on the Rise of Online Hate Speech. [online
- Wired. (2020). Toxicity in Gaming Is Dangerous. Here's How to Stand Up to It. [online

- Wu, S., Lin, T.-C. and Shih, J.-F. (2017) “Examining the antecedents of online disinhibition,” Information technology & people, 30(1), pp. 189–209.
- Yano, A. (2020). The Australian government holds Facebook and Google accountable. [online
- Yardi, S. and Boyd, D. (2010) Dynamic debates: An analysis of group polarization over Time on twitter, Umich.edu. Available at: https://yardi.people.si.umich.edu/pubs/Yardi_DynamicDebates.pdf (Accessed: September 17, 2021).
- Yin, W. and Zubiaga, A. (2021) “Towards generalisable hate speech detection: a review on obstacles and solutions,” PeerJ. Computer science, 7(e598), p. e598.
- Yin, W. and Zubiaga, A. (2021) “Towards generalisable hate speech detection: a review on obstacles and solutions,” PeerJ. Computer science, 7(e598), p. e598.
- York, J.C. (2021). The delights and the dangers of deplatforming extremists. [online
- Zampieri, M. et al. (2019) “Predicting the type and target of offensive posts in social media,” in Proceedings of the 2019 Conference of the North. Stroudsburg, PA, USA: Association for Computational Linguistics.
- Zannettou, S. et al. (2017) “The Web centipede: Understanding how Web communities influence each other through the lens of mainstream and alternative news sources,” arXiv [cs.SI
- Zero Fox (2019) What is domain protection and how to address domain-based attacks (2019) Zerofox.com. Available at: https://www.zerofox.com/blog/domain-protection-top-3-domain-based-attack-tactics-and-how-to-address-them/ (Accessed: September 17, 2021).