European Journal on Criminal Policy and Research (2023) 29:477–493
https://doi.org/10.1007/s10610-023-09543-z

'Is This a Hate Speech?' The Difficulty in Combating Radicalisation in Coded Communications on Social media Platforms

Benjamin Farrand[1]

Accepted: 26 April 2023 / Published online: 5 May 2023

Abstract
The use of the Internet for the purposes of radicalisation is well understood. The use of social media platforms to spread messages of hate and intolerance has become mainstreamed, both as a means of communication and a focus of academic attention. The purpose of this paper is to focus on the complications in addressing these radicalisation efforts where the means of communication is through Internet memes, in which humour and coded language are used as means of radicalising an audience. While existing legal frameworks provide for online platforms to tackle this problem through a combination of assuring immunity from liability for taking action when content is brought to a platforms attention, as well as engaging in voluntary proactive measures, this paper explores the difficulty of addressing content that is more borderline in nature and where arguments concerning humour and freedom of expression may be raised by those spreading these messages in a system providing for significant discretion on the part of online platforms. Considering developments under the Digital Services Act, and an increased focusing on the algorithmic control of content, this article will argue that even these new measures may find the countering of radicalising content conveyed through humour quite difficult.

Keywords Hate speech · Online regulation · Platforms · Freedom of expression · Memes · Digital Services Act

Introduction

At its very core, the Internet is about communication, and innovations in the digital communication infrastructure have resulted in substantial increases in audience. According to Statista, from 2017 to 2021, the number of social media users has increased from 2.73 billion to 4.26 billion (Statista, 2022). Yet as the quantity of users and the volume of messages have risen, so too have the associated regulatory challenges. From the spread of copyright infringing material on streaming platforms (Frosio, 2017), hate speech aimed at

✉ Benjamin Farrand
ben.farrand@ncl.ac.uk

[1] Newcastle Law School, 19-24 Windsor Terrace, Newcastle University, Newcastle upon Tyne NE1 7RU, UK

diverse religious and racial groups (Vidgen & Yasseri, 2020), women and LGBT communities (Ging & Siapera, 2018), and the global proliferation of political and health-related disinformation (Carrapico & Farrand, 2021; Hameleers et al., 2020), new means of communication typified by immediacy and wide reach have resulted in the unparalleled dissemination of forms of speech that are considered offensive, extreme, or socially undesirable (Nilan, 2021). At the centre of regulatory endeavours seeking to control the spread of such messages are the social media companies that provide the platform for these communications to reach their target audiences. The Digital Services Act, Regulation 2022/2065, is the EU's latest initiative at controlling the dissemination of illegal forms of content, placing obligations upon online platforms to improve their processes and procedures for identifying and mitigating the impacts of this content on their services. Yet how successful is this new initiative likely to be?

The purpose of this article is to explore the EU's current attempts to regulate content on social media by way of a model in which decisions regarding the managing of content on online platforms are taken by the platform operators. It does this through exploring a case study on the difficulties in combating hate-focused radicalisation efforts expressed through coded communications. The title of this article, *Is This a Hate Speech?*, refers to the 'Is this a pigeon?' meme, in which a slightly confused young scholar (which is, fittingly, an android) misidentifies a butterfly. Much like the android confusing the butterfly with a pigeon due to their capacity for flight, the systems put in place to control hate speech online pose the risk of misidentification where hate and non-hate messages share humorous framing, but with more serious potential outcomes. This article argues that the regulation of hate speech at the EU level prioritises control of this content by service providers as the result of its initial regulatory decisions, resulting in platforms themselves taking the role of arbiter, determining whether communications made using their services constitute hate speech, or are communications which, while offensive, nevertheless benefit from protection based on the principle of freedom of expression.

An approach of minimal intervention and self-regulation, in which private sector operators of platforms are tasked with the identification, determination, and removal of alleged hate speech-coded communications along with the promotion of voluntary codes of conduct, has resulted in a form of policy 'lock-in'. New legislative initiatives such as DSA, which impose obligations on online platforms regarding the operation of their services in tackling illegal content, are ultimately the result of these previous decisions. However, the considerable discretion afforded to these platforms means that combating hate on their services that are expressed through coded communication such as memes becomes very difficult to achieve, as they occupy a space of 'non-obvious' hate, making proactive regulation of such content controversial and subject to assessments of intent.

It must be clearly stated here that the purpose of this article is *not* to debate the contours of freedom of expression in the context of offensive speech, which as discussed in the next section has been covered extensively in the work of other academics. Nor is it intended to consider the national interpretation and application of hate speech laws, which is beyond the scope of this article. Instead, it is to explore how the EU's regulatory decision-making concerning online platforms as regulated self-regulators has impacted upon legislative initiatives such as the DSA, which in this context making tackling certain forms of hate online exceptionally difficult. In analysing the EU's policymaking in this area, the main documentary sources are legislative instruments and codes of conduct focused on the enforcement of content moderation rules by online, coupled with relevant policy documents associated with such initiatives. It does not, therefore, cover every piece of content-related legislation seeking to harmonise national laws, but those placing specific moderation obligations upon

private sector intermediaries to demonstrate trajectories of law-making relating to private sector enforcement in this domain and how it results in a system ill-equipped for dealing with certain forms of coded communication.

'I Hate You. Just Kidding, But Not Really': Hate Speech and Radicalisation Through Coded Communication

'Hate speech' is a fundamentally contested concept (Titley et al., 2014, p.10). For some scholars, the concept is fundamentally flawed, both in its linkage of speech to harm and in terms of being widespread enough to warrant regulation (Bennett, 2016). Others argue that 'viewpoint-discriminatory' laws intended to prevent hate speech actually result in increased intolerance (Weinstein, 2017). For supporters of hate speech regulation, there is sufficient evidence to link speech to harm, particularly when those speech acts incite or promote violence against certain groups (Peršak, 2022). Others still argue that the protection of human dignity compels us to prohibit abusive conduct such as hate speech that could threaten it (see Brown, 2015). Waldron in particular has made the case for considering the harms associated with hate speech as impacting upon dignity and safety from violence to the extent that regulation is both necessary and welcome (Waldron, 2012).

What becomes immediately clear is that hate speech is an inherently subjective (Kabasakal Badamchi, 2021; Kumaresan & Vidanage, 2019), culturally dependent (Boromisza-Habashi, 2012), and socially contextual (Cowan & Hodge, 1996) phenomena in communication. As such, it is also highly normatively charged debate, with proponents and opponents of the regulation of speech often finding little in the way of common ground (an excellent overview of the competing normative positions can be found in Billingham & Bonotti, 2019). However, for the purposes of this article, which is not to debate the line between protection from harm and freedom of expression, but to assess regulatory responses to hate speech, we can nevertheless provide a working definition drawing from institutional responses to the phenomena. The Council of Europe, for example, says that hate speech should be understood as:

The advocacy, promotion, or incitement, in any form, of the denigration, hatred or vilification of a person or group of persons, as well as any harassment, insult, negative stereotyping, stigmatization or threat in respect of such a person or group and the justification of all the preceding types of expression, on the ground of 'race', colour, descent, national or ethnic origin, age, disability, language, religion or belief, sex, gender, gender identity, sexual orientation and other personal characteristics or status. (Council of Europe Commission against Racism and Intolerance, 2015, p.3)

Hate speech has the potential to radicalise. Derogatory language about immigrants and ethnic minority communities, for example, has been shown to radicalise those exposed to it, decreasing levels of empathy with marginalised groups, increasing the normalisation of 'in group and out group' attitudes, and desensitising those exposed to hate speech to the extent it no longer presents as extreme or offensive in nature (Bilewicz & Soral, 2020). Through this process of radicalisation, individuals may be more likely to believe that violence or discrimination against those groups is justified in order to achieve political goals, such as the (perceived) security, purity, or homogeneity of a particular population (Dal Santo & D'Angelo, 2022). In the context of the Internet, research has predominantly focused on radicalisation by Islamist groups (for an overview see Kadivar, 2017; Joshi,

2021), with an increased focus on extreme right groups, including the 'very online' alt-right (Boatman, 2019; Gray, 2018; Hawley, 2017; Holt et al., 2015).

The alt-right has been described as 'a creature of the Internet, where many of its members, even some of the most prominent, are anonymous or tweet under pseudonyms […] it's a movement with several factions which shrink or swell according to the political breeze and the task at hand' (Wendling, 2018, p.5). 'Members' of this movement may hold white nationalist, antisemitic, Islamophobic, misogynistic or anti-LGBT views, or a combination of these views (a comprehensive account of the different ideological groupings can be found in Hermansson et al., 2020). These groups often refer to themselves as adopting scientific or rational perspectives, juxtaposed against a 'woke' majority that is both derided and perceived to constitute an ideological threat (Finlayson, 2021). A commonality between the alt-right and Islamist groups such as Daesh is in the radicalisation processes — both groups have operated through creating in-group acceptance, support, and validation of existing beliefs, reinforcing the sense of superiority that the group feels towards those populations they consider as inferior and/or threatening to them (on radicalisation by Daesh, see Murshed & Pavan, 2011; on the alt-right see Boatman, 2019). Once those community links have been formed, more extreme ideological elements are introduced to new members of that community, which either encourage directly violent action or support for violent action where it has taken place. This creates what has been defined as 'stochastic terror', defined as 'the use of mass media to provoke random acts of ideologically motivated violence that are statistically predictable but individually unpredictable' (Hamm & Spaaij, 2017, p.84).

This is key in understanding the difficulties in tackling forms of hate speech online. Coded communication has been a studied feature of racist speech since the 1970s, for example, with the use of symbols and behaviours as ways of derogating out-groups (McConahay & Hough Jr., 1976), with terms such as 'welfare queens' used to denigrate Black women beginning in particular communities (Macedo & Bartolomé, 1999), which are then mainstreamed by political and media actors, normalising what should be considered extreme speech (Kelly, 2010). Humour is one such way of coding language around hatred; the Ku Klux Klan, for example, has been argued to use jokes as a means of conveying racist stereotypes and promote violence while being able to claim that the intent is to be humorous rather than to spread hate (Billig, 2001). This type of coding is particularly common online, where messages can be conveyed not only through textual jokes, but also using memes, gifs, and other forms of mixed media to communicate ideas to new audiences or reinforce beliefs within existing communities.

The use of the Internet has allowed extremist groups to gain access to less radical audiences, using humour as a means of communicating their ideas to this new audience, facilitating engagement with radicalising messaging that those individuals may not have been exposed to offline (Ekman, 2014; May & Feldman, 2018). The use of humour allows hate speech to hide 'in plain sight', through ironic misdirection, presenting exaggeratedly distorted racist or misogynistic stereotypes that *could* potentially be seen as satirical but *could* equally be the position held by the author of the text or image. One such example is the Finspång meme, in which Swedish far right individuals share images of people hung from cranes with the message 'see you in Finspång', a place designated as the town where national traitors will be executed (Centre for Analysis of the Radical Right, 2018). According to Askanius, the adaptations of this meme entered into Swedish mainstream culture, representing a blending of serious messaging regarding the murder of political traitors with surreal or comical imagery, creating a 'hate–humour nexus' that allows for widening the discursive space of acceptable political discourse (Askanius, 2021, p.152). Similarly,

the viral spread of the Pepe the frog meme, adopted as a symbol by alt-right individuals (against the wishes of the original author), used to spread racist and antisemitic messages from the position of underdogs or victims of those targeted groups (Glitsos & Hall, 2019).

Research by Woods and Ruscher suggests that the far-reaching spread of memes, originating on fringe or extremist websites before appearing on large mainstream platforms like Facebook, results in an air of tolerance for these more extreme forms of humour while also being effective at recruiting new members to those political causes or radicalising them into further action (Woods and Ruscher, 2021). However, for those not immersed in those cultures, understanding the subtexts or hate conveyed in the ambiguous messaging is difficult, which can pose significant problems for the effective regulation of these forms of radicalising speech. Through these humorous ambiguities, 'the codes and symbols may be deployed either to avoid detection or to enhance the mystery, and mystique, around the nature of the alt-right' (Miller-Idriss, 2018, pp.123–127). In the next section of the article, we will explore these difficulties further, beginning with an assessment of how the initial regulatory decisions taken in the EU have led to a distinct approach to governing online speech. This approach in turn has limited the range of potential approaches to managing content on social media platforms, with consequences for effectively countering hate spread through coded communications.

Understanding Content Moderation by Platform Operators Through Historical Institutionalism

Content moderation, defined for the purposes of this article as the process of screening, evaluating, and approving or suppressing communications by users of an online platform (De Gregorio, 2020; Flew et al., 2019; Zeng & Kaye, 2022), has been largely dictated in the EU by an initial approach of regulated self-regulation, which has then shaped future interventions. These developments can be traced using historical institutionalism. Historical institutionalism identifies the ways in which institutions are formed, evolve, and structure themselves (Fioretos et al., 2018), with institutions constituting 'the formal or informal procedures, routines, norms and conventions embedded in the organisational structure of the polity' (Hall & Taylor, 1996). In the context of this article, the institutions focused upon constitute the rules and organisations (on this see Streeck & Thelen, 2005) dictating the EU's approach to content regulation. Decisions taken and the rules thereby implemented have a historical legacy that determines the scope of what future rules are deemed appropriate and legitimate within a given sphere of activity. What appear to be small or simple choices at the time can therefore have significant long-term impacts (Sorensen, 2015). These structuring decisions serve to facilitate some actions, while restricting others, and are known as 'path dependences' (Steinmo et al., 1992). Path dependence means that 'each step down a particular pathway increases the likelihood of further steps along the same pathway, and increases the cost of reverting to some previously available option' (Sorensen, 2015, p.21).

Of relevance to this article are the concepts of layering and conversion, both mechanisms by which policy change can be affected without having serious disruptions to existing path dependencies. Layering is the process which involves 'the grafting of new elements onto an otherwise stable institutional framework' (Thelen, 2004, p.35). As will be discussed, this includes placing new responsibilities or requirements on platforms, which do not significantly alter the regulatory landscape. Conversion entails adopting new goals

or frameworks to which the existing ruleset or approach is applied as actors redeploy the rules in a way that suits their interests as they apply them (Ertugal, 2021). Rules from one policy sector will be 'converted' and applied in a new policy sector, such as taking rules applied to copyright enforcement and applying them to restricting hate speech. Its contribution to this article is in showing the underlying ideas that influenced the approach to the regulation of platforms in this domain and then how those initial regulatory decisions regarding the role of platforms in moderating content then influenced subsequent legislation such as the DSA.

Self-regulation by private sector operators is not a new or unusual phenomenon and has been discussed by authors such as Ogus (1995), who considered it may be based in economic rationales concerning costs and efficiencies arising from the technical knowledge and expertise possessed by actors within those sectors (Ogus, 1994, pp.110–111; see also Elkin-Koren & Salzberger, 2004). With the move to the regulatory state, in which the state sets frameworks for rules that are then implemented by private actors (Majone, 1994, 1997; Yeung, 2010), re-regulation of particular sectors through institutional dynamics in which the state is active in steering policy direction while regulatory agencies and business actors do the rowing (Levi-Faur, 2005) gradually has become typified by more 'networked' regulatory structures (Black, 2001; Coen & Thatcher, 2008), in which the private sector is involved in not only 'rowing' but also 'steering', with industry best standards and practices being used as the basis for obligations laid down in regulatory regimes (see, for example, Carrapico & Farrand, 2017). This can result in the establishment of regulatory regimes in which sectors are subject to 'regulated self-regulation', in which the private sector actors devise the rules by which they are scrutinised, which has been increasingly the case when considering content moderation relating to hate speech.

In this context, the EU rules laid down in the E-Commerce Directive (2000/31/EC) were devised in an environment in which regulated self-regulation was actively promoted as regulating in areas typified by private sector infrastructure ownership and deference to the technical expertise of those private sector actors (Christou & Simpson, 2004). The subsequent policy decisions expanding the content moderation approach in the EU are the result of path dependencies originating in these rules, with new developments being in the form of gradual changes enacted through layering and conversion. The position of the European Commission was that new digital technology companies should be regulated in such a way as to guarantee that market activity flourished and that European economies were able to take advantage of these new developments, rather than stifle them through excessive regulation (Farrand, 2023). In the communication preceding the directive, a preference was demonstrated for self-regulatory codes — 'any legislative action should impose the fewest possible burdens on the market' (European Commission, 1997, p.14), indicating the ideas regarding regulation driven by economic efficiencies as considered by Ogus as cited above.

As a result of this 'minimalist intervention' approach, the intermediary immunity from liability provisions encoded in Articles 12–14 of the E-Commerce Directive borrowed heavily from the principles of the US's Digital Millennium Copyright Act (see, for example, McEvedy, 2002; Peguera, 2008), as well as s.230 of the Communications Decency Act (Edwards, 2018). As has been argued by Husovec, rules concerning the liability of intermediaries can been characterised as 'accountability without liability' (Husovec, 2017). Under Article 14, Internet service providers hosting content would not be considered liable for the content or actions of their users if that content was deemed to be illegal (or infringing upon copyright), so long as the service provider acted expeditiously to remove that content once it was brought to their attention (this is an exhaustively covered topic in the literature, but some key sources on this include Julià-Barceló & Koelman, 2000; Rizzuto,

2012). Furthermore, there would be no general obligation to monitor the usage of services provided under Article 15. The first cases concerning the interpretation and application of these principles originated in disputes over the protection of copyright online. In *Scarlet v SABAM* (C-70/10), the Court of Justice of the European Union (CJEU) made it clear that while the protection of copyright was a key value of the European Union, required under the Information Society Directive (2001/39/EC) and its fundamental rights obligations, so too was the protection of freedom of expression under Article 10 of the European Convention of Human Rights (ECHR) and Article 11 of the EU's Charter of Fundamental Rights (EUCFR), and any measures taken to remove content online must ensure that these values were respected (para.115 of the decision). This freedom of expression is framed as freedom of information, entailing both the right to distribute and to receive it (see, for example, Geiger et al., 2020).

At the level of general rules, then, the origins of platform governance of content have been based on minimalist interventions and self-regulation, with guidance that freedom of expression is to be given due regard in the context of removals of content (Husovec & Peguera, 2015; Hoboken & Keller, 2019). This places online service providers including social media platforms in a difficult position as both enforcers of rules regarding content and arbiters of human rights protections (see, for example, Jørgensen & Pedersen, 2017). As was stated in the introduction to this article, the purpose of this article is not to determine whether the balance between the protection of freedom of expression and protection from harm has been correctly decided, but to consider the way in which the initial rule formulations have dictated the direction of policy and what this means for platforms dealing with complex cases of radicalising speech imparted through humour and memes. In this respect, taking a historical institutionalist perspective, the original rules have acted as a basis for a self-regulatory approach being taken by online service providers, providing significant flexibility and discretion, while obligations to respect fundamental rights have been layered over the top of those rules, with elements of conversion as principles determined in copyright-related cases at the CJEU have been broadened to cover content regulation generally by these service providers. In the next section, we will look at the development of regulatory initiatives concerned with hate speech specifically, demonstrating the path dependence that has arisen from the original rule structure.

From Immunity from Liability to a Sense of Responsibility? Changing Perceptions of the Role of Platforms

To understand the approach taken to platform governance in the context of radicalising hate speech, it is useful to refer to the EU's legal framework, which has influenced the direction that social media platforms have taken in content moderation. The Framework Decision on combating certain forms and expressions of racism and xenophobia by means of criminal law (2008/913/JHA) governs the EU's current approach to hate speech. This legislation requires Member States to ensure criminal prosecution for public incitement or hatred directed against groups or members of groups on the basis of race, colour, religion, descent, or national or ethnic origin, or for actions such as holocaust denial (2008/913/JHA, Article 1). This includes communication through the distribution of tracts, pictures, or other material. Within this approach, however, Member States are obliged to ensure that fundamental rights to freedom of expression are protected under Article 7. Platforms have increasingly been brought into the regulatory structures for combating hate speech online,

aligned the networked governance approach discussed in the previous section. Under a Code of Conduct introduced by the Commission in partnership with Facebook, Twitter, YouTube, and Microsoft in May 2016, online platforms agreed to voluntarily introduce measures to combat hate speech as defined in the 2008 Framework Decision (European Commission, 2016a). These measures, designed to complement the existing terms and conditions and best practices of the platforms, included measures such as providing clear and effective review processes, develop notice and flagging systems and educational/awareness-raising campaigns (European Commission, 2016a, pp.2–3). All this would be done, however, while recognising 'the need to defend the right to freedom of expression' (European Commission, 2016a, p.1). In terms of policy formulation, the approach taken here is not reflective of critical juncture, but of layering, as voluntary commitments to ensure self-regulatory practices were implemented in line with the requirements of existing legislation were layered over the original approach taken in the E-Commerce Directive. This constitutes reinforcement of the regulated self-regulation approach, where it is the terms of service and guidelines of the platforms that serve as the basis for decisions regarding the moderation of content, and what communications are to be removed or deprioritised, and which are to be left up.

The Code of Conduct supplemented the immunity from liability for removal of content brought to a platform's attention with the encouragement of voluntary proactiveness. A preliminary report on the effectiveness of the Code of Conduct on Illegal Hate Speech found that in the first months of its operation, out of the 600 notifications made (including 270 by 'trusted flaggers'), only 28.2% of content was removed, but with Twitter and YouTube, removal was much more likely if notification was made by a trusted flagger (European Commission, 2016b, p.4). In 2017, the Commission released a Communication on tackling illegal content online, where it observed that despite efforts aimed at reducing such content, the spread was both wider and increasing in speed and that 'online platforms […carry] a significant societal responsibility in terms of protecting users and society at large and preventing criminals and other persons involved in infringing activities online from exploiting their services' (European Commission, 2017, p.2). While recognising the important role of platforms, the Commission nevertheless did not propose significant regulatory changes that would either impact upon the immunity from liability regime or go beyond voluntary proactive measures.

It is important to note that this marked an important development in the discourse concerning the role of platforms in governing their services, which was also clear from the language used around platforms in the context of disinformation and hybrid security threats in 2016 (European Commission and High Representative of the Union for Foreign Affairs and Security Policy, 2016) and the belief that greater accountability was required on the part of platforms that were increasingly considered as contributing to security threats in the EU, rather than being more 'neutral' providers of economically beneficial services (see, for example, Carrapico & Farrand, 2020). However, in terms of *policies* as distinct from *rhetoric*, the new Communication did not represent any significant rupture, but instead a continuation of existing policies, with layering of recommendations on how to remove content considered illegal online within the context of platform self-regulation, with measures including aiming to remove content such as hate speech and terrorism-related content within 24 h. Within this, however, was a requirement that providers do not 'over-remove' content, which was perceived as impinging upon freedom of expression (European Commission, 2017, p.6). Platforms were encouraged to ensure that there was transparency in notice and takedown proceedings, as well as safeguards put in place to prevent over-blocking, thereby ensuring protection for freedom of expression (European Commission, 2017, pp.16–17). In this respect, flexibility and discretion

were maintained on the part of platforms, provided with a framework for a goal to be achieved, but left to determine the best means to achieve those goals themselves.

Consideration of how platforms approach these issues can be evidenced by the Terms of Service or Community Guidelines of the platforms included in the initial Code of Conduct. According to the first available version of Facebook's Community Guidelines from May 2018, terms that could result in removal included tier 1 offences such as threats of violence against members of a protected group or dehumanising speech or imagery and tier 2 offences including statements or imagery conveying inferiority, or the use of expressions such as 'I don't like' or 'I hate' (Meta, 2018). In a March 2019 update to these guidelines, however, the company stated that 'we allow humour and social commentary related to these topics' (Meta, 2019). This in turn was reiterated in the June 2020 update with the policy rationale for the hate speech guidelines (Meta, 2020). In comparison, assessment of Meta's Facebook Community Guidelines regarding incitement to violence (which includes its policies on terrorism) does not make such references to humour (Meta, 2022a). In announcing their rules on reducing hateful conduct, Twitter made it clear in 2017 that it applied not only to promoting violence or abuse based on protected characteristics, but also the use of hateful imagery, including logos, symbols, or images (Twitter Safety, 2017). A 2019 update stated that while Twitter encouraged people to express themselves, this did not extend to abuse, and they therefore prohibited 'language that dehumanizes others on the basis of religion, caste, age, disability, disease, race, ethnicity, national origin, gender, gender identity, or sexual orientation', providing example of proscribed content (Twitter Safety, 2019). Unlike the Facebook Community Guidelines, however, Twitter did not explicitly refer to humour in its discussion of these policies. These initiatives have been successful in tackling the 'low hanging fruit' of hate speech in these fields, where hate is clearly and directly expressed, as evidenced by more recent reports on the effectiveness of the Codes of Conduct. As the 2019 report clearly states (in the view of the European Commission, and seemingly reinforcing the path dependence of original regulatory decisions), 'self-regulation works' (European Commission, 2019, p.1).

While this was constructed as being a victory for protection of fundamental rights, given the discussion of the conveying of radicalising messages of hate through humour and indirect means in the second section of this article, this may also potentially indicate that the inherent tensions in protecting freedom of expression while tackling hate speech result in a certain level of policy incoherence that legal and self-regulatory systems are not easily able to reconcile. In this respect, here, we see indications of policy drift — the old rules of the 2000s era Internet of the E-Commerce Directive being applied to an infinitely more complicated and variable Internet of the 2010s (Hoboken & Keller, 2019). As more direct messages of hate became more likely to be removed from these services, then more indirect and coded means of communicating those messages became more likely to be used to convey the same sentiments. Given the proclivities of Internet users, this entailed an increased reliance on in-jokes, meta-humour, and subversiveness that make it incredibly difficult to distinguish between humour intended to satirise and humour intending to appear to satirise while hiding its underlying message in plain sight. This, as well as the difficulties for platforms in combating this approach, will be considered in the next section.

The Digital Services Act: Bringing Platforms into the Regulatory Space

The von der Leyen Commission has sought to be more assertive in the regulation of online activities in order to protect the EU's 'digital sovereignty', from controlling political advertising and disinformation to combating hate speech and radicalisation online

(European Commission, 2020b; see also Farrand & Carrapico, 2022). Amid concerns that social media platforms were not sufficiently engaged in preventing the spread of malicious communications (Carrapico & Farrand, 2020), the Commission issued a proposal for a regulation clarifying the obligations for social media platforms, the Digital Services Act (European Commission, 2020a). The proposal stated that it was to build upon measures such as the Code of Conduct on illegal hate speech (2020a, p.5), introducing a horizontal framework intending to define the rules defining the responsibilities and obligations of digital service providers, 'and online platforms in particular' (2020a, p.1). It does not create new rules regarding illegal content including hate speech, but instead seeks to clarify how the platforms are to deal with this illegal content on their services. Here, the path dependence leading from the original regulatory decisions made in the late 1990s is visible; the Commission is explicit that the DSA Act builds 'on the key principles set out in the E-Commerce Directive, which remain valid today' (2020a, p.2). It states the intention of the Regulation will be to contribute to online safety while protecting fundamental rights and maintains the liability provisions (including Article 14) from the E-Commerce Directive (2020a, pp.2–3).

Where the DSA goes further than the existing framework is through policy layering. In the final version of the Regulation (2022/2265), the intermediary immunity from liability provisions is maintained in Articles 4–8. Of interest is Article 9, which concerns orders to act against illegal content. Within a new 'co-regulatory structure', orders to remove content can be made by national authorities to online service providers, who can then provide information to the national authorities and by contacting Digital Services Coordinators in a Member State, who can then relay information to the other Digital Servicer Coordinators in the other Member States. Under Article 34, very large online platforms such as Facebook and Twitter are required to perform risk assessments including both an assessment of the risks posed by illegal content on their services and impacts on fundamental freedoms such as freedom of expression. Article 35 requires that mitigation efforts are made to minimise the impacts of those risks. The Digital Services Act is incredibly comprehensive in setting out the processes expected of service providers including online platforms (Cauffman & Goanta, 2021), but following previous regulatory interventions, leaves it to the platforms themselves to determine how best to fulfil the obligations imposed (Farrand, 2023; Maelen, 2022). Oversight is provided, indicating a less 'minimal intervention' model, insofar as the risk assessments and mitigation strategies are subject to external scrutiny, but they are still based upon best practices of commercial operators in those fields, as well as their own terms of service and risk assessments. Even such oversight as provided is relatively light touch, given the ability of the very large online platforms to choose their own external auditor (Laux et al., 2021)

What do these changes mean for the tackling of hate speech conveyed through memes and other forms of coded communication? In essence, the Digital Services Act does not fundamentally change the approach — the Code of Conduct on Illegal Hate Speech remains the basis for actions in this field (albeit widened to include a larger number of platforms such as TikTok), and upon the basis of Framework Decision 2008/913/JHA. It is worth noting that while the EU has determined that it may be appropriate to expand the list of hate speech offences to include hate speech on the basis of gender, gender identity, and sexual orientation (European Commission, 2021), it is to the credit of the larger platforms such as Twitter and Facebook that they have already included these types of hate speech into their community guidelines and have made considerable efforts in removing misogynistic and homophobic content. What *has* changed, however, is a growing recognition of the scale of the 'ironic hate' problem online on the part of policymakers. According

to a report published by the European Commission's Radicalisation Awareness Network, 'humour has become a central weapon of extremist movements to subvert open societies and to lower the threshold towards violence […] it rebrands extremist positions in an ironic guise, blurring the lines between mischief and potentially radicalising messages' (Fielitz & Ahmed, 2021, p.4). Communities on fringe message boards such as 4chan or subreddits on reddit produce memes that contain messages of hate, which can then go viral and spread to larger platforms such as Facebook (Rauf, 2021), which means that Meta's risk assessment would have to factor in contagion from smaller platforms *not* categorised as very large online platforms and not subject to the same scrutiny and oversight. It is in this context that the Commission has introduced Regulation 2021/784 on the dissemination of terrorist content online, which under Article 3 requires content hosts to remove or disable access to terrorist content within 1 h of receiving the removal order from a national authority. One particular group that the Commission has expressed concern about as constituting a potential terrorist threat are 'incels' who become radicalised through online coded communications, with crossover with other more extremist violent groups such as the alt-right in the form of 'stormcels', 'whitecels', or 'alt-rightcels' (European Commission & Radicalisation Awareness Network, 2021, p.8), who make an interesting case study of the difficulties in tackling this content. Incels consider themselves as victims of the 'feminisation' of society and see women as paradoxically superior and unattainable while also being inferior with an obligation to serve men (C. R. Kelly & Aunspach, 2020). The group is characterised by a perception of victimisation and ostracisation and susceptible to self-radicalisation (Daly & Reed, 2022). This has resulted in terrorist attacks, such as those committed by a self-identified incel in Canada in 2018, and another in Plymouth in 2021. These groups are particularly notorious for the use of humour as a means of covertly coding extremist messages (European Commission & Radicalisation Awareness Network, 2021, p.16).

An example of this is the subversion of the 'virgin vs Chad' meme (representing 'low confidence' and 'high confidence' characters, respectively), used as a way of developing identity while expressing ant-feminist messages (Aulia & Rosida, 2022; Lindsay, 2022). However, the virgin vs Chad meme is also used to humorously critique incel culture and has also been adopted to cover a wide range of activities, beliefs, and culture that go far beyond male sexuality (such as 'the virgin Roman vs the Chad Carthaginian' that appears to be based predominantly around attitudes towards the use of shield walls). This makes identifying when these messages are being spread to radicalise and espouse anti-women discourse exceptionally difficult. Similarly, 'Pepe the frog', a frequent symbol in alt-right messaging, has also been used by Hong Kong-based democracy protestors either unaware of the coded meaning of the image or as a form of re-adoption and changed meaning (Peters & Allan, 2022). Given the volume of communications, platforms are increasingly relying upon algorithms to make decisions regarding content removal. Concerns have been expressed regarding algorithms over-removing content with negative impacts upon freedom of expression (see, for example, Dias Oliva, 2020; Senftleben, 2020), with writers such as Elkin-Koren describing them as a 'black box' that makes the underlying logic of the code used inscrutable and lacking in transparency (Elkin-Koren, 2012; Perel & Elkin-Koren, 2015). Ethical concerns regarding the use of algorithms have also been raised, including how decisions are made (Yeung, 2019) and the potential for discriminatory effects and threats to autonomy (Danaher, 2019). 'Under-removal' is also a potential problem. As Fuchs and Schäfer state, 'implicit forms of abuse pose difficulties […] ironic usages of language […] which can also be meant to be abusive, are not only particularly difficult to detect for machine learning processes or sentiment analysis, but are often even hard to grasp for the human researcher' (Fuchs and Schäfer, 2021, p.556). The use

of subtlety and irony can be used to avoid detection both by algorithms and human moderators not familiar with a particular usage, making detection much more difficult (Bhat & Klein, 2020). 'Saying the quiet part out loud' is an expression that refers to someone being explicit in their messaging and is generally frowned upon as plausible deniability is lost — for these communities, the reliance upon irony, meta-humour, and memes makes identification of hate speech much more difficult, both through automated content moderation and by trusted flaggers or other human intermediaries. As Meta's latest version of its Community Standards on Hate Speech state, 'in certain cases, we will allow content that may otherwise violate the Community Standards when it is determined that the content is satirical' (Meta, 2022b). Yet how can Meta determine what is satire expressing critique of its message, and what is ironic 'satire' intended in hiding a message in plain sight? Particularly where such decisions are taken on the basis of algorithms in the first instance, and given concerns over Meta's approach to the prioritisation and (dubious) deprioritisation of potentially harmful content (Cyphert & Martin, 2022; Zenone et al., 2023), concerns must be raised over the significant levels of discretion and flexibility afforded by the regulated self-regulatory model. With Elon Musk's takeover of Twitter, it has been announced that its content moderation guidelines will change, although it has not been announced in which form, in line with Musk's views regarding freedom of speech that are self-described as 'absolutist' (Davies, 2022). Significant cuts to the content moderation team, along with a very negative assessment of Twitter's current safeguards in the first report on the functioning of the enhanced code of conduct disinformation issues by the Commission (Goujard, 2023), suggest these issues may well get worse.

Conclusions

The regulation of speech online is a complicated one. Leaving aside the issue of balancing protection from harm with freedom of expression, even should a regulator wish to actively tackle hate speech online, doing so in practice is incredibly difficult. Part of this difficulty relates to the myriad of ways that hate is expressed, some so self-referential and meta in nature that understanding outside a given community is low to non-existent. But another part of this difficulty relates to path dependencies and regulatory choices. The E-Commerce Directive set up a system of self-regulation with minimal state intervention, on the basis that commerce should be encouraged to flourish online. Furthermore, by reinforcing the importance of freedom of expression as a factor to consider in all moderation platforms yet leaving it to the platforms themselves to make these decisions, certain types of approach to content regulation became more possible, while other options were potentially restricted. Further developments have continued along the lines of leaving the content decisions to the platforms, on the basis of a regulatory system determined on principles of immunity from liability in the late 1990s, which continue to be reiterated as the right approach today. Instead, we have seen policy layering and conversion, both through the encouragement of voluntary codes of conduct and oversight and compliance mechanisms intended to scrutinise the decisions made by those platforms, rather than dictate what those decisions should be. However, the current approach struggles when dealing with hate that is conveyed through 'non-traditional' means, such as meta-humour, irony, and memes, which can conceal meaning and intent while nevertheless radicalising particular audiences. Attempts to combat this type of content are likely to struggle, both in cases of algorithmic control and human intervention. And while these messages of hate are conveyed in the

form of humour, which is reinforced as important in the context of freedom of expression, regulatory solutions based on regulated self-regulation by platforms will struggle further, given the significant discretion and flexibility afforded to them in tackling these issues. In this context, it would appear the current system for content moderation cannot meme.

Acknowledgements The author would like to sincerely thank Melanie McLaughlan for her excellent research assistance in the process of formulating this article. The author would also like to thank the anonymous reviewers of this paper for their insights, questions, and recommendations for refining the article.

Data Availability Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

Declarations

Conflict of Interest The author declares no competing interests.

References

Askanius, T. (2021). On frogs, monkeys, and execution memes: Exploring the humor-hate nexus at the intersection of neo-Nazi and alt-right movements in Sweden. *Television & New Media, 22*(2), 147–165. https://doi.org/10.1177/1527476420982234

Aulia, M. P., & Rosida, I. (2022). The phenomenon of involuntary celibates (incels) in Internet meme culture: A reflection of masculine domination. *International Journal of Media and Information. Literacy, 7*(1) Article 1.

Bennett, J. T. (2016). The harm in hate speech: A critique of the empirical and legal bases of hate speech regulation. *Hastings Constitutional Law Quarterly, 43*(3), 445–536.

Bhat, P., & Klein, O. (2020). Covert hate speech: White nationalists and dog whistle communication on Twitter. In G. Bouvier & J. E. Rosenbaum (Eds.), *Twitter, the Public Sphere, and the Chaos of Online Deliberation* (pp. 151–172). Springer International Publishing. https://doi.org/10.1007/978-3-030-41421-4_7

Bilewicz, M., & Soral, W. (2020). Hate speech epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology, 41*(S1), 3–33. https://doi.org/10.1111/pops.12670

Billig, M. (2001). Humour and hatred: The racist jokes of the Ku Klux Klan. *Discourse & Society, 12*(3), 267–289. https://doi.org/10.1177/0957926501012003001

Billingham, P., & Bonotti, M. (2019). Introduction: Hate, offence and free speech in a changing world. *Ethical Theory and Moral Practice, 22*(3), 531–537. https://doi.org/10.1007/s10677-019-10027-5

Black, J. (2001). Decentring regulation: Understanding the role of regulation and self regulation in a "post-regulatory" world. *Current Legal Problems, 54*(1), 103–146.

Boatman, E. (2019). The kids are alt-right: How media and the law enable white supremacist groups to recruit and radicalize emotionally vulnerable individuals. *Law Journal for Social Justice, 12*(Fall), 2–61.

Boromisza-Habashi, D. (2012). The cultural foundations of denials of hate speech in Hungarian broadcast talk. *Discourse & Communication, 6*(1), 3–20. https://doi.org/10.1177/1750481311427793

Brown, A. (2015). *Hate speech law: A philosophical examination*. Taylor & Francis https://library.oapen.org/handle/20.500.12657/25902

Carrapico, H., & Farrand, B. (2017). 'Dialogue, partnership and empowerment for network and information security': The changing role of the private sector from objects of regulation to regulation shapers. *Crime, Law and Social Change, 67*(3), 245–263.
Carrapico, H., & Farrand, B. (2020). Discursive continuity and change in the time of Covid-19: The case of EU cybersecurity policy. *Journal of European Integration, 42*(8), 1111–1126. https://doi.org/10.1080/07036337.2020.1853122
Carrapico, H., & Farrand, B. (2021). When trust fades, Facebook is no longer a friend: Shifting privatisation dynamics in the context of cybersecurity as a result of disinformation, populism and political uncertainty. *JCMS. Journal of Common Market Studies, 59*(5), 1160–1176.
Cauffman, C., & Goanta, C. (2021). A new order: The digital services act and consumer protection. *European Journal of Risk Regulation, 12*(4), 758–774. https://doi.org/10.1017/err.2021.8
Centre for Analysis of the Radical Right. (2018). 'Finspång' – An execution meme of the Swedish radical right ignites the political discourse. *CARR Blog* https://www.radicalrightanalysis.com/2018/07/06/finspang-an-execution-meme-of-the-swedish-radical-right-ignites-the-political-discourse/
Christou, G., & Simpson, S. (2004). Emerging patterns of E-commerce governance in Europe: The European Union's directive on E-commerce. 32nd Telecommunications Policy Research Conference: Communication, Information and Internet Policy. George Mason University Law School, Arlington, Virginia, U.S., October 1-3, 2004, 1–35
Coen, D., & Thatcher, M. (2008). Network governance and multi-level delegation: European networks of regulatory agencies. *Journal of Public Policy, 28*(1), 49–71. https://doi.org/10.1017/S0143814X08000779
Council of Europe Commission against Racism and Intolerance (2015). ECRI General Policy Recommendation No 15 on Combating Hate Speech (CRI(2016)15)
Cowan, G., & Hodge, C. (1996). Judgments of hate speech: The effects of target group, publicness, and behavioral responses of the target. *Journal of Applied Social Psychology, 26*(4), 355–374. https://doi.org/10.1111/j.1559-1816.1996.tb01854.x
Cyphert, A., & Martin, J. (2022). "A change is gonna come:" Developing a liability framework for social media algorithmic amplification. *UC Irvine Law Review, 13*(1), 155.
Dal Santo, E., & D'Angelo, E. (2022). Relationship of online hate, radicalization, and terrorism. In E. W. Dunbar (Ed.), *Indoctrination to Hate: Recruitment Techniques of Hate Groups and How to Stop Them*. ABC-CLIO.
Daly, S. E. & Reed, S. M. (2022). "I think most of society hates us": A qualitative thematic analysis of interviews with incels. *Sex Roles*, 86(1-2), 14–33. https://doi.org/10.1007/s11199-021-01250-5
Danaher, J. (2019). The ethics of algorithmic outsourcing in everyday life. In K. Yeung & M. Lodge (Eds.), *Algorithmic Regulation* (pp. 91–118). Oxford University Press. https://doi.org/10.1093/oso/9780198838494.003.0005
Davies, P. (2022). *Musk's Twitter: A platform for free speech or extremist hate?* Euronews https://www.euronews.com/next/2022/10/28/will-elon-musk-s-twitter-become-a-beacon-of-free-speech-or-a-soap-box-for-hate-speech
De Gregorio, G. (2020). Democratising online content moderation: A constitutional framework. *Computer Law & Security Review, 36*, 105374. https://doi.org/10.1016/j.clsr.2019.105374
Dias Oliva, T. (2020). Content moderation technologies: Applying human rights standards to protect freedom of expression. *Human Rights Law Review, 20*(4), 607–640. https://doi.org/10.1093/hrlr/ngaa032
Edwards, L. (2018). "With great power comes great responsibility?": The rise of platforms liability. In L. Edwards (Ed.), *Law, Policy and the Internet* (pp. 253–290). Hart.
Ekman, M. (2014). The dark side of online activism: Swedish right-wing extremist video activism on YouTube. *MedieKultur: Journal of Media and Communication Research, 30*(56), 56. https://doi.org/10.7146/mediekultur.v30i56.8967
Elkin-Koren, N. (2012). Governing access to user-generated content: The changing nature of private ordering in digital networks. In *Governance%2C Regulation and Powers on the Internet*. Cambridge University Press. https://doi.org/10.1017/CBO9781139004145.020
Elkin-Koren, N., & Salzberger, E. M. (2004). *Law, economics and cyberspace: The effects of cyberspace on the economic analysis of law*. Edward Elgar.
Ertugal, E. (2021). Hidden phases of de-Europeanization: Insights from historical institutionalism. *Journal of European Integration, 43*(7), 841–857. https://doi.org/10.1080/07036337.2020.1869955
European Commission. (1997). A European initiative in electronic commerce. *COM, 97*(157), 1–27.
European Commission. (2016a). *Code of conduct on countering illegal hate speech online*
European Commission. (2016b). Code of conduct on countering illegal hate speech online: First results on implementaton (pp. 1–4)
European Commission. (2017). Tackling illegal content online: Towards an enhanced responsibility of online platforms. *COM, 2017*(555).

European Commission. (2019). Code of conduct on countering illegal hate speech online: Fourth evaluation confirms self-regulation works (pp. 1–6)
European Commission. (2020a). Proposal for a regulation on a single market for digital services (Digital Services Act) and amending Directive 2000/31/EC. *COM, 2020*(825), 1–112.
European Commission. (2020b). Shaping Europe's digital future (pp. 1–9)
European Commission. (2021). A more inclusive and protective Europe: Extending the list of EU crimes to hate speech and hate crime. *COM, 2021*(777).
European Commission & Radicalisation Awareness Network. (2021). *Incels: A first scan of the phenomenon (in the EU) and its relevance and challenges for P/CVE* (pp. 1–21). European Commission.
European Commission and High Representative of the Union for Foreign Affairs and Security Policy. (2016). Joint framework on countering hybrid threats. *JOIN, 2016*(18), 1–18.
Farrand, B. (2023). The Ordoliberal Internet? Continuity and change in the EU's approach to the governance of cyberspace. *European Law Open, 2*(1), 1–40.
Farrand, B., & Carrapico, H. (2022). Digital sovereignty and taking back control: From regulatory capitalism to regulatory mercantilism in EU cybersecurity. *European Security, 31*(3), 435–453. https://doi.org/10.1080/09662839.2022.2102896
Fielitz, M., & Ahmed, R. (2021). It's not funny anymore. In *Far-right extremists' use of humour* (pp. 1–18). European Commission, Radicalisation Awareness Network.
Finlayson, A. (2021). Neoliberalism, the alt-right and the intellectual dark web. *Theory, Culture & Society, 38*(6), 167–190. https://doi.org/10.1177/02632764211036731
Fioretos, O., Falleti, T. G., & Sheingate, A. (2018). Historical institutionalism in political science. In O. Fioretos, T. G. Falleti, & A. Sheingate (Eds.), *The Oxford Handbook of Historical Institutionalism*. Oxford University press.
Flew, T., Martin, F., & Suzor, N. (2019). Internet regulation as media policy: Rethinking the question of digital communication platform governance. *Journal of Digital Media & Policy, 10*(1), 33–50. https://doi.org/10.1386/jdmp.10.1.33_1
Frosio, G. (2017). From horizontal to vertical: An intermediary liability earthquake in Europe. *Journal of Intellectual Property Law & Practice, 12*(7), 565–575. https://doi.org/10.1093/jiplp/jpx061
Fuchs, T., & Schäfer, F. (2021). Normalizing misogyny: Hate speech and verbal abuse of female politicians on Japanese Twitter. *Japan Forum, 33*(4), 553–579. https://doi.org/10.1080/09555803.2019.1687564
Geiger, C., Frosio, G., & Izyumenko, E. (2020). Intermediary liability and fundamental rights. In G. Frosio (Ed.), *Oxford Handbook of Online Intermediary Liability*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198837138.013.7
Ging, D., & Siapera, E. (2018). Special issue on online misogyny. *Feminist Media Studies, 18*(4), 515–524. https://doi.org/10.1080/14680777.2018.1447345
Glitsos, L., & Hall, J. (2019). The Pepe the Frog meme: An examination of social, political, and cultural implications through the tradition of the Darwinian Absurd. *Journal for Cultural Research, 23*(4), 381–395. https://doi.org/10.1080/14797585.2019.1713443
Goujard, C. (2023). *Elon Musk's Twitter fails first EU disinformation test*. POLITICO https://www.politico.eu/article/elon-musk-twitter-fails-eu-first-disinformation-test-digital-services-act/
Gray, P. W. (2018). 'The fire rises': Identity, the alt-right and intersectionality. *Journal of Political Ideologies, 23*(2), 141–156. https://doi.org/10.1080/13569317.2018.1451228
Hall, P. A., & Taylor, R. C. R. (1996). Political science and the three new institutionalisms. *Political Studies, 44*(5), 936–957. https://doi.org/10.1111/j.1467-9248.1996.tb00343.x
Hameleers, M., Powell, T. E., Meer, T. G. L. A. V. D., & Bos, L. (2020). A picture paints a thousand lies? The effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media. *Political Communication, 37*(2), 281–301. https://doi.org/10.1080/10584609.2019.1674979
Hamm, M. S., & Spaaij, R. (2017). *The age of lone wolf terrorism*. Columbia University Press.
Hawley, G. (2017). *Making sense of the alt-right*. Columbia University Press.
Hermansson, P., Lawrence, D., Mulhall, J., & Murdoch, S. (2020). *The international alt-right: Fascism for the 21st century?* Routledge. https://doi.org/10.4324/9780429032486
Hoboken, J. V., & Keller, D. (2019). Design principles for intermediary liability laws. *Transatlantic Working Group, IViR* (University of Amsterdam). 1–11
Holt, T., Freilich, J. D., Chermak, S., & McCauley, C. (2015). Political radicalization on the Internet: Extremist content, government control, and the power of victim and jihad videos. *Dynamics of Asymmetric Conflict, 8*(2), 107–120. https://doi.org/10.1080/17467586.2015.1065101
Husovec, M. (2017). *Injunctions against intermediaries in the European Union: Accountable but not liable?* Cambridge University Press.
Husovec, M., & Peguera, M. (2015). Much ado about little—Privately litigated internet disconnection injunctions. *International Review of Intellectual Property and Competition Law, 46*(1), 10–37.

Jørgensen, R. F., & Pedersen, A. M. (2017). Online service providers as human rights arbiters. In M. Taddeo & L. Floridi (Eds.), *The Responsibilities of Online Service Providers* (pp. 179–199). Springer International Publishing. https://doi.org/10.1007/978-3-319-47852-4_10

Joshi, R. (2021). Religious radicalization in France: Contextualizing the 2021 'Anti-Separatism' Bill. *Strategic Analysis*, 1–6. https://doi.org/10.1080/09700161.2021.1966870

Julià-Barceló, R., & Koelman, K. J. (2000). Intermediary liability: Intermediary liability in the e-commerce directive: So far so good, but it's not enough. *Computer Law & Security Review, 16*(4), 231–239. https://doi.org/10.1016/S0267-3649(00)89129-3

Kabasakal Badamchi, D. (2021). Hate speech and limits of free speech. In D. Kabasakal Badamchi (Ed.), *Dimensions of free speech: An exploration of a new theoretical framework* (pp. 119–156). Springer International Publishing. https://doi.org/10.1007/978-3-030-88319-5_7

Kadivar, J. (2017). Online radicalization and social media: A case study of Daesh. *International Journal of Digital Television, 8*(3), 403–422. https://doi.org/10.1386/jdtv.8.3.403_1

Kelly, M. (2010). Regulating the reproduction and mothering of poor women: The controlling image of the welfare mother in television news coverage of welfare reform. *Journal of Poverty, 14*(1), 76–96. https://doi.org/10.1080/10875540903489447

Kelly, C. R., & Aunspach, C. (2020). Incels compulsory sexuality and fascist masculinity. *Feminist Formations* 32(3), 145–172. https://doi.org/10.1353/ff.2020.0044

Kumaresan, K., & Vidanage, K. (2019). HateSense: Tackling ambiguity in hate speech detection. *National Information Technology Conference (NITC), 2019*, 20–26. https://doi.org/10.1109/NITC48475.2019.9114528

Laux, J., Wachter, S., & Mittelstadt, B. (2021). Taming the few: Platform regulation, independent audits, and the risks of capture created by the DMA and DSA. *Computer Law & Security Review, 43*, 105613. https://doi.org/10.1016/j.clsr.2021.105613

Levi-Faur, D. (2005). The rise of regulatory capitalism: The global diffusion of a new order. *The ANNALS of the American Academy of Political and Social Science, 598*(1), 12–32. https://doi.org/10.1177/0002716204272590

Lindsay, A. (2022). Swallowing the black pill: Involuntary celibates' (Incels) anti feminism within digital society. *International Journal for Crime, Justice and Social Democracy, 11*(1) Article 1. https://doi.org/10.5204/ijcjsd.2138

Macedo, D., & Bartolomé, L. I. (1999). Dancing with bigotry. In D. Macedo & L. I. Bartolomé (Eds.), *Dancing with bigotry: Beyond the politics of tolerance* (pp. 1–33). Palgrave Macmillan US. https://doi.org/10.1007/978-1-137-10952-1_1

Maelen, C. V. (2022). Hardly law or hard law? Investigating the dimensions of functionality and legalisation of codes of conduct in recent EU legislation and the normative repercussions thereof. *European Law Review, 47*(6), 752–772.

Majone, G. (1994). The rise of the regulatory state in Europe. *West European Politics, 17*(3), 77–101. https://doi.org/10.1080/01402389408425031

Majone, G. (1997). From the positive to the regulatory state: Causes and consequences of changes in the mode of governance. *Journal of Public Policy, 17*(2), 139–167.

May, R., & Feldman, M. (2018). Understanding the alt-right. Ideologues, 'Lulz' and Hiding in Plain Sight. In *Understanding the alt-right. ideologues, 'Lulz' and Hiding in Plain Sight* (pp. 25–36). Verlag. https://doi.org/10.1515/9783839446706-002

McConahay, J. B., & Hough, J. C., Jr. (1976). Symbolic racism. *Journal of Social Issues, 32*(2), 23–45. https://doi.org/10.1111/j.1540-4560.1976.tb02493.x

McEvedy, V. (2002). The DMCA and the ECommerce Directive. *European Intellectual Property Review, 24*(2), 65–73.

Meta. (2018). *Hate speech | Transparency Centre*. Meta Community Guidelines https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/

Meta. (2019). *Hate speech | Transparency Centre*. Meta Community Guidelines https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/

Meta. (2020). *Hate speech | Transparency Centre*. Meta Community Guidelines https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/

Meta. (2022a). *Violence and incitement | Transparency Centre*. Facebook Community Guidelines https://transparency.fb.com/en-gb/policies/community-standards/violence-incitement/

Meta. (2022b). *Hate speech | Transparency Centre*. Meta Community Guidelines https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/

Miller-Idriss, C. (2018). What makes a symbol far right? Co-opted and missed meanings in far-right iconograpy. In M. Fielitz & N. Thurston (Eds.), *Post-Digital Cultures of the Far Right: Online Actions and Offline Consequences in Europe and the US*. transcript Verlag.

Murshed, S. M., & Pavan, S. (2011). Identity and Islamic radicalization in Western Europe. *Civil Wars, 13*(3), 259–279. https://doi.org/10.1080/13698249.2011.600000
Nilan, P. (2021). Online discourse and social media. In P. Nilan (Ed.), *Young People and the Far Right* (pp. 29–56). Springer. https://doi.org/10.1007/978-981-16-1811-6_2
Ogus, A. (1994). *Regulation: Legal form and economic theory*. Hart Publishing.
Ogus, A. (1995). Rethinking self-regulation. *Oxford Journal of Legal Studies, 15*(1), 97–108.
Peguera, M. (2008). The DMCA safe harbors and their European counterparts: A comparative analysis of some common problems. *Columbia Journal of Law and the Arts, 32*(4), 481–512.
Perel, M., & Elkin-Koren, N. (2015). Accountability in algorithmic copyright enforcement. *Stanford Technology Law Review, 19*, 473.
Peršak, N. (2022). Criminalising hate crime and hate speech at EU level: Extending the list of Euro-crimes under Article 83(1) TFEU. *Criminal Law Forum, 33*(2), 85–119. https://doi.org/10.1007/s10609-022-09440-w
Peters, C., & Allan, S. (2022). Weaponizing memes: The journalistic mediation of visual politicization. *Digital Journalism, 10*(2), 217–229. https://doi.org/10.1080/21670811.2021.1903958
Rauf, A. A. (2021). New moralities for new media? Assessing the role of social media in acts of terror and providing points of deliberation for business ethics. *Journal of Business Ethics, 170*(2), 229–251. https://doi.org/10.1007/s10551-020-04635-w
Rizzuto, F. (2012). The liability of online intermediary service providers for infringements of intellectual property rights. *Computer and Telecommunications Law Review, 18*(1), 4–15.
Senftleben, M. (2020). The original sin – Content 'moderation' (censorship) in the EU. *GRUR International, 69*(4), 339–340. https://doi.org/10.1093/grurint/ikaa025
Sorensen, A. (2015). Taking path dependence seriously: An historical institutionalist research agenda in planning history. *Planning Perspectives, 30*(1), 17–38. https://doi.org/10.1080/02665433.2013.874299
Statista. (2022). *Number of worldwide social network users from 2018 to 2027*. Statista https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/
Steinmo, S., Thelen, K., & Longstreth, F. (1992). *Structuring politics: Historical institutionalism in comparative analysis*. Cambridge University Press.
Streeck, W., & Thelen, K. A. (2005). *Beyond continuity: Institutional change in advanced political economies*. Oxford University Press.
Thelen, K. (2004). *How institutions evolve: The political economy of skills in Germany, Britain, the United States, and Japan*. Cambridge University Press.
Titley, G., Keen, E., & Földi, L. (2014). *Starting points for combating hate speech online* (pp. 1–92). Council of Europe Youth Department.
Twitter Safety. (2017). *Enforcing new rules to reduce hateful conduct and abusive behavior*. Twitter Safety Blog https://blog.twitter.com/en_us/topics/company/2017/safetypoliciesdec2017
Twitter Safety. (2019). *Updating our rules against hateful conduct*. Twitter Safety Blog https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate
Vidgen, B., & Yasseri, T. (2020). Detecting weak and strong Islamophobic hate speech on social media. *Journal of Information Technology & Politics, 17*(1), 66–78. https://doi.org/10.1080/19331681.2019.1702607
Waldron, J. (2012). The harm in hate speech. In *The Harm in Hate Speech*. Harvard University Press. https://doi.org/10.4159/harvard.9780674065086
Weinstein, J. (2017). Hate speech bans, democracy, and political legitimacy symposium: Hate speech and political legitimacy. *Constitutional Commentary, 32*(3), 527–584.
Wendling, M. (2018). *Alt-right: From 4chan to the White House* (1st ed.). Pluto Press.
Woods, F. A., & Ruscher, J. B. (2021). Viral sticks, virtual stones: Addressing anonymous hate speech online. *Patterns of Prejudice, 55*(3), 265–289. https://doi.org/10.1080/0031322X.2021.1968586
Yeung, K. (2010). The regulatory state. In R. Baldwin, M. Cave, & M. Lodge (Eds.), *The Oxford Handbook of Regulation*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199560219.003.0004
Yeung, K. (2019). Why worry about decision-making by machine? In K. Yeung & M. Lodge (Eds.), *Algorithmic Regulation* (pp. 21–48). Oxford University Press. https://doi.org/10.1093/oso/9780198838494.003.0002
Zeng, J., & Kaye, D. B. V. (2022). From content moderation to visibility moderation: A case study of platform governance on TikTok. *Policy & Internet, 14*(1), 79–95. https://doi.org/10.1002/poi3.287
Zenone, M., Kenworthy, N., & Maani, N. (2023). The social media industry as a commercial determinant of health. *International Journal of Health Policy and Management, 12*(1), 1–4. https://doi.org/10.34172/ijhpm.2022.6840

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH ("Springer Nature").

Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users ("Users"), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use ("Terms"). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.

These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.

We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.

While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.

These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.

To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.

Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.

If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com