Open Library of Humanities Journal

Young, J 2025 Dogwhistles, Discrimination, Humour and the Law: Regulating Implicit Messaging. *Open Library of Humanities*, 11(2): pp. 1–26. DOI: https://doi.org/10.16995/olh.19789

Open Library of Humanities

Dogwhistles, Discrimination, Humour and the Law: Regulating Implicit Messaging

Jennifer Young, University of Groningen, The Netherlands; Honorary Associate, Department of Law and Social Justice, University of Liverpool, United Kingdom, subridere@gmail.com

This paper explores how implicit, discriminatory messages bypass sanctions in the United Kingdom and beyond, despite their potential for significant societal harm. Drawing on linguistic and humour research, it emphasises the role of humour used to conceal discriminatory expression and evade legal boundaries. The study extends understanding of how courts and, more recently, online platform moderators sometimes struggle to identify discrimination in humour, especially satire and irony. This has raised concerns about potential regulatory overreach as well as the risk that failing to address the issue could normalise hateful expression. Expanding laws to regulate implicit speech is problematic; it risks suppressing legitimate expression and stifling creativity. Nonetheless, implicit expressions used to promote social division or discrimination are equally problematic if left unchallenged. Therefore, tackling this issue requires a multi-faceted approach, combining education about the legal consequences of both offline and online expression, platform policies, and media literacy initiatives. These initiatives should help audiences better interpret implicit messages, complemented by legal literacy to encourage consideration of the legal implications of their speech.

Introduction

This article cites examples of discriminatory and derogatory humour in political or social discourse, referring to language that appears harmless or innocuous on the surface, but which carries a hidden, divisive, or discriminatory message to a specific audience. This is commonly known as the practice of dogwhistling. Provided here is, therefore, an explanation of what dogwhistles are and how humour can be used to increase their dissemination whilst further masking any explicit hate speech.

Dogwhistles exploit the legal grey areas surrounding freedom of expression by using implicit language to incite hostility. Combined with humour, their opacity allows speakers to deny discriminatory meanings in a plausible way (Ajder and Glick, 2023; Elder, 2024). In such a context, humour is defined as '[u]tterances which are identified [...] on the basis of paralinguistic, prosodic and discursive clues, as intended by the speaker(s) to be amusing and perceived to be amusing by at least some of the participants' (Holmes and Marra, 2002: 67).

Adjacent to this issue are international hate speech legal thresholds and how these are balanced against protecting freedom of expression. The article thus examines legal cases involving discriminatory speech, humour and freedom of expression and compares examples of court cases, media regulation and Meta Oversight Board decisions. The final section considers the hard and soft law approaches to identifying implicit discriminatory humour and what steps might be taken to minimise negative effects on society. It is necessary to address these phenomena because dogwhistles are effective recruitment tools for the far right to deepen divisions and damage democracy (Fielitz and Reem, 2021).

Currently, there is no single solution which might effectively counter the challenges of socio-political dogwhistles that do not reach the legal threshold of hate speech or incitement, meaning the law could sanction the speaker.[1] Many scholars and governments, and bodies such as the UN and the European Council, have concluded that this issue requires a multi-faceted approach. This combines a better understanding of dogwhistling tactics with a more effective decoding of implicit language within platform policies and literacy initiatives. These approaches should include a legal literacy approach for speakers to understand where the legal limit to expression lies, explaining the potential for punitive repercussions, so they might reconsider their words.

Why Dogwhistle?

The term 'dogwhistle' gained traction in the 1980s in the United States to describe covert political messaging. The expression is used because these messages behave like

[1] Throughout this article the term 'speaker' includes the creator or publisher of audio/visual media and 'speech' includes audio/visual output such as memes and cartoons.

a conventional dog whistle; by using implicit language, symbols or codes, only the intended audience can hear the call. They contain an overt message which some of the audience understand at its anodyne face value and a concealed discriminatory message targeted at a subset of the audience (Guercio and Caso, 2023: 4). Dogwhistles are often used by politicians to express opinions which might be unacceptable to some voters whilst others would find these same opinions appealing (Henderson and McCready 2018: 231). By using implicit messaging, instead of an overtly controversial or offensive message, the political actor can deny any incriminating meaning (Sayeed et al., 2024).

Dogwhistles can manipulate people 'in ways that they would resist if the manipulation was carried out more openly—often drawing on racist attitudes that are consciously rejected' (Saul, 2024: 9). Furthermore, the engaging guise of humour can mitigate and normalise stereotypes and inequality where they should be considered unspeakable (Tsakona, 2024: 124). Other groups such as the far right use these tactics with the aim of undermining social cohesion, 'othering' certain vulnerable or minority groups and creating a chilling effect on participatory democracy (Boholm and Sayeed, 2023: 53). The European Union highlights the need to help young people identify dogwhistles (Fielitz and Reem, 2021: 12). Moreover, humour discourages an audience from critically considering the content and so can further enable tolerance for discriminatory views (Tsakona, 2024: 106).

According to Jennifer Saul, there are two forms of discriminatory dogwhistles. The first is the Overt Code where they 'work like a secret code, designed to be understood by one group (those who don't accept norms against racism) and not recognized by the others' (2024: 9). It is directed at a particular audience, primed by a prior awareness of the speaker's political or ideological position, who can decode the encoded language. Saul's second form of racist dogwhistle is the Covert Effect. Here, people who would ordinarily be anti-racist are manipulated by racialized imagery rather than racial terms. Many examples in the literature are drawn from the United States, one being 'welfare queen', implicitly reinforcing the myth that Black women want handouts and are lazy (Saul, 2024: 9). Using comic incongruity, it suggests that one can live like a queen on welfare, combating the reality of those drawing social security as being in poverty. Such terminology causes ordinary people who would usually adhere to anti-racist norms 'to base their decisions on racial resentments' (Saul, 2024: 9). These terms do not include offensive, explicit discriminatory language, they simply imply meanings devised to incite hatred and social division on a subconscious level (Fogal, Harris and Moss, 2018: 548-549). Dogwhistles buried in such texts send out a signal which says to the audience, 'you are not alone'; and, by doing so, they act as a recruitment call and embolden real-world action offline (Kasimov, Johnston and Heer, 2023) and targeted harassment online (Marwick, 2021).

These techniques often evade detection by online moderation algorithms. It is not easy for platforms such as Meta, Google and X (formerly Twitter) to tackle user-generated content which uses coded or implicit language, especially when it takes the form of humour. Although X, now it is owned by Elon Musk, seems to be actively promoting divisive and contentious speech, and its policy on hateful conduct (X's policy on hateful conduct, 2023) often seems to be little more than lip service (Hutchinson, 2024). Even where other platforms have more stringent policies on hate speech, such as Meta, who established an Oversight Board to supervise its decisions, implicit content is problematic.

There is a distinction too between offline and online content regulatory or moderation approaches. Online expression sits in an 'interstitial zone', and the different regulatory regimes are spread across the sectors governed by political, private and civil societies. It is considered by some as 'ungovernable' because it is decentralised, swift in message dissemination, and its coded language can 'flout law and regulation'. The modus operandi of constantly making new words, for example the word 'rapefugees', ensures that algorithmic and human content moderators do not immediately recognise such words as hate speech (Ganesh, 2018; Baider, 2022). Some would find this play on words amusing, and the use of humour to both disguise and amplify discriminatory expression is an effective way to evade sanctions. The meaning of 'rapefugee' is obvious to most people, perpetuating the myth of refugees not as vulnerable humans but rather as being sexually aggressive criminals. By the time the content moderators detect and understand these new words, the term will have spread both online and offline and been promoted by their target audience. The content producers will have moved on to inventing other forms of expression to raise laughs and evade moderation.

It is not just new words which make detection problematic for automated moderators. Established brand names are used as codes; examples include 'terms like "Google", "Skittle", and "Yahoo" as substitutes for offensive words describing Black people, Muslims, and Mexicans' (Kantrowitz, 2016). Without context, the practice of content moderation by AI and automation cannot differentiate between what is innocent and what is a coded racist slur. The coded nature means it is unlikely to reach a threshold where it might be sanctioned by traditional media laws as incitement to hatred or violence. As Gillespie (2020) has pointed out, AI content moderation is not necessarily the answer to regulating social media and could possibly even have counter effects on the outcomes pursued by platforms.

In using implicit language, the speaker has another layer of protection and can plausibly deny that their speech is a call to hostility or that they intended anything legally contentious. If people are manipulated by implicit language, this makes such

speech more dangerous than explicit hate speech, as in many jurisdictions, the United States being a notable outlier, hate speech can be legitimately restricted. When explicit language is used, then the speaker's intention to incite hatred, for example, is easier to detect, and intent is an important consideration when applying sanctions. Implicit messaging can fall into that grey area short of criminality (GIFCT, 2023), as will be illustrated with case law below. Whilst implicit language is more subtle, the effect and the target are often the same as explicit speech.

Plausible Deniability through Humour

Humour can give speech an additional protection from legal sanctions on at least two fronts, one being that satire is often afforded a higher level of protection from State restrictions (Godioli, Young and Fiori, 2022). The second defence is that humour provides the speaker with an opportunity to plausibly deny hateful intent (Elder, 2024). A speaker can claim satire, or irony, or that it was just a joke and that there was no intent to stir up hostility (Matamoros-Fernández and Jude, 2025). When judging if satirical speech can be legitimately restricted the European Court of Human Rights defines satire as

> a form of artistic expression and social commentary and, by its inherent features of exaggeration and distortion of reality, naturally aims to provoke and agitate [Para. 33].[2]

Therefore, using humour as a front to discriminatory expression increases the difficulty of holding a speaker to account and of restricting the speech legitimately without State authorities appearing overly censorious. The Court states that 'any interference with an artist's right to such expression must be examined with particular care'[para.33].[3]

Parody has been defined by the Court of Justice of the European Union as expression which must 'evoke an existing work while being noticeably different from it' and 'constitute an expression of humour or mockery' [para.20].[4] However, exact meaning for terms including parody and satire remain 'nebulous' (Jacques, 2019: 1). Like satire and parody, irony similarly may be used to mock either a vulnerable target or the act of discrimination itself, and the audience may need to read between the lines. In other words, such genres require a close reading and in such cases a discourse analysis may be necessary to understand the target of the speech (Simpson, 2023: 114). Therefore, the task for the courts, broadcasters and online moderators is often a difficult one, to ascertain if the contested satirical or ironic material is promoting or critiquing socially

[2] *Vereinigung Bildender Künstler v. Austria* (68354/01 – 25th January 2007).

[3] *Vereinigung Bildender Künstler v. Austria* (68354/01 – 25th January 2007).

[4] Judgment in Case C-201/13 *Johan Deckmyn and Vrijheidsfonds VZW v Helena Vandersteen and Others.*

problematic material. The differentiation is important because humour can function to divide social groups and reinforce social boundaries (Pérez, 2017). A veneer of humour has the additional benefit of making content popular for sharing whilst minimising the serious nature of the harm hidden in the speech (Zinigrad, 2024).

A report by the British media regulator The Office of Communications (Ofcom) explains how humour encourages people to read and share harmful conspiracy theories, misinformation and disinformation (Strong, Owen and Mansfield, 2023). One reason for this is that humour is a very effective medium to spread discriminatory ideology both off and online. Mathilda Åkerlund (2021) has pointed out how 'hateful humour' is used 'to express hateful ideas without explicitly stating them'. This humorous veneer means 'jokes', or what is alleged to pass as a funny jibe or mocking turn of phrase, can be shared unwittingly. This results in stock phrases being repeated and perpetuated by those who do not recognise the hidden message.

When it comes to hateful expression online or offline, humorous or otherwise, the International Covenant on Civil and Political Rights (ICCPR) Article 20 (United Nations Office of the High Commissioner for Human Rights, 1966) and the American Convention on Human Rights Article 13 (American Convention on Human Rights, 1969) require an advocacy of hatred for the State to legitimately restrict it. This advocacy has been interpreted as an intent requirement (Mendel, 2010), defined as 'an intentional and public promotion of hatred; the advocated 'hatred' is supposed to constitute incitement to discrimination, hostility or violence, i.e. illegal material actions' (Bayer, 2021). Yet intent is difficult to establish when that message is implicit rather than explicit, and more so when framed as humour because of plausible deniability. Furthermore, this can then lead to a claim of victimhood by the speaker, as they can argue their words were just misunderstood, or that the audience does not have a sense of humour.

In 2021, the European Commission published *It's not funny anymore. Far-right extremists' use of humour* (Fielitz and Ahmed, 2021). The paper explains how the far-right has embraced the tool of humour to lower people's thresholds towards discriminatory and violent content. Such groups have learned that a successful movement should be 'entertaining and participatory', and a shared sense of humour is a way to immerse people 'into extremist ideologies'. One example for this dates back to 2017, when the online news site *Huffpost* ran an article about a 'style guide' produced by white supremacist website *The Daily Stormer*. This told their writers not to be explicit in their intention to be discriminatory, but rather,

> when using racial slurs, it should come across as half-joking – like a racist joke that everyone laughs at because it's true. [...] The reader is at first drawn in by curiosity

> or the naughty humor, and is slowly awakened to reality by repeatedly reading the same points (Feinberg, 2017).

It has been suggested that the amplifying effect of humour should be considered an important factor in assessing hate speech cases (Zinigrad, 2024). Hate speech causes societal damage, negatively affects social cohesion and undermines civic values of equality and participation in public life (Gov.UK, n.d.), but further legal restrictions affect the fundamental right to freedom of expression. There are ways to restrict content without turning to the law. For example, *The Daily Stormer* was disconnected from various online infrastructure companies and its domain refused registration in different countries until eventually, a security and distribution network cancelled its contracts. It was forced to move to the 'dark web', which made the content difficult to find. Again, though, such moves by private companies are controversial because of the implications on the regulation of free expression and the norms of due process (Suzor, 2019: 3-9).

Balancing Hate Speech and Freedom of Speech

Whilst hate speech is not protected in most democratic jurisdictions, with the United States being a noticeable exception with its near absolutist approach to freedom of expression, any legitimate State restriction on speech must be balanced with the right to freedom of expression. In democratic countries, there are robust speech protections, especially regarding political speech (Rowbottom, 2012). These are protected under both national and supranational Courts – the African Charter on Human and Peoples' Rights, the American Convention on Human Rights and the European Convention on Human Rights, as well as the ICCPR. However, there is no unqualified right to freedom of expression. Article 20 (2) of the ICCPR gives permissible limitations on freedom of expression, including the prohibition of speech which advocates 'national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence'. This gives some protection to those who are targeted by hateful speech. These limitations on speech are judged with a three-part test, the restriction must be provided for by law, in pursuit of a legitimate aim, and necessary and proportionate.

The Office of the United Nations High Commissioner for Human Rights (OHCHR), the agency which promotes and protects human rights, advocates for a high threshold when assessing the legitimacy of speech restrictions. This is reflected within their guidelines, the Rabat Plan of Action (on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence).[5]

[5] The Rabat Plan of Action (on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence) (A/HRC/22/17).

This is a six-part threshold test which needs to be reached for expression to amount to a criminal offence (United Nations Office of the High Commissioner for Human Rights, 2020). The Plan has been used by many states, including some national authorities for audio-visual communication, the European Court of Human Rights, and private bodies such as Facebook's (Meta's) Oversight Board. The tests include the social and political context, status of the speaker, intent to incite the audience against a target group and the content and form of the speech.

The status of the speaker is considered especially important. Figures of authority, especially 'high-level politicians', are more able to appeal to an audience and therefore their speech is believed to have a greater impact (McKevitt, 2019). The Plan states that 'Political and religious leaders should refrain from using any incitement to hatred', but a later example illustrates that this guidance is ignored by populist politicians.

As with Article 19 of the ICCPR, the intent of the speaker is paramount to the Rabat test. Here lies a problem when dogwhistle tactics and humour mask the intent of the speaker. The Rabat Plan considers that intent

> requires the activation of a triangular relationship between the object and subject of the speech act as well as the audience (OHCHR, n.d.).

This triangular relationship reflects the complexity of speech reception. It refers to the object and subject of the speech (potentially a vulnerable target) and the audience. If sections of the audience know the speaker and share their political ideology, then they would recognise any implicit meaning pitched towards incitement which others might miss (Bandrowski, 2024). Whilst the Rabat Plan of Action is useful in judging hate speech acts, as its test aims for a 'multidimensional contextualisation of messages [...] necessary to understand and to assess covert hate speech' (Baider, 2022: 2354), it does not fully address the issue. It can set the threshold too high to sanction implicit messages, even when those could affect social cohesion and democratic processes.

Populist politicians use dogwhistle tactics to hide hateful rhetoric, masking xenophobia and discriminatory language to gain political advantage. Satire and comedy are used by authoritarian and populist actors to transform the public sphere, and now humour is also a tool of the powerful and not simply one of resistance against those in power (Beck and Spencer 2025).

Populism and Promoting Division, a Regulatory Conundrum

In the lead-up to the United Kingdom's 2024 general election, some British politicians were using divisive slogans such as 'I want my country back'. Following Brexit, it echoed

the UK Independence Party's slogan 'I want my sovereignty back' (Baker and Scott, 2024). If we decode this kind of language, consider the speaker (here a white, male, right-wing MP), and the political party he had recently defected to (Reform UK, an anti-immigration party), the words could carry another message, namely 'I want to take our country back from the foreigners who have come over here with their different religions and cultures changing our traditional British way of life'. Or, more succinctly, some might understand it as, 'let's get Britain back to being white and ethnically British'.

Such slogans appeal to people's emotions. 'We want our country back' was a refrain repeated during recent protesting and rioting in the UK, so this rhetoric is recognised by the target audience and leads to serious, real-life consequences. Nearly five hundred people were arrested, and it was reported that the violence was 'fuelled by misinformation online, the far right and anti-immigration sentiment' (Kotecha, 2024). In one of the cases, the accused claimed in an interview that he was using 'dark humour', but this was dismissed by the judge. He was sentenced for publishing written material with the intention of stirring up racial hatred (Mistry, 2024).

Billig described similar rhetoric by a British Prime Minister more than 30 years ago, explaining how the repetitive use of the 'patriotic card' has, 'amongst its rhetorical potentialities, that familiar monster: the self-righteous call to national anger' (1995: 93-94). It is divisive and hostile, it denigrates diversity and inclusion, whilst in the natural and ordinary meaning of the words, and to a reasonable person, means nothing. In such circumstances, plausible deniability protects the speaker from being accused of race-baiting or inciting hostility. Using implicit language gives the speaker 'a theoretical "get out" in the face of linguistic adversity' (Elder, 2024). Additionally, it is more effective than racist or explicit anti-immigration rhetoric in gaining support for xenophobic policies because of the covert effects (Sayeed et al., 2024). These apparently meaningless messages and slogans are then repeated by the mainstream media in news reporting of the politicians' speeches (Lewis and Marwick, 2017). Being repeated by reputable news agencies gives them further gravitas. Politicians have become adept at repeating these slogans when being interviewed by the media, who amplify, reinforce and normalise this rhetoric. It is unlikely that 'I want my country back' would pass the legal threshold tests, even though the status of the speaker means he should be held to higher accountability.

Authorities and courts need to navigate through complex issues of freedom of expression, consider the speaker, intent and accountability, and not least decide which of the multiple meanings in an implicit communication should take precedence. To illustrate this, what follows are examples of the varying approaches to implicit language addressed by the courts and the British broadcast regulator Ofcom in a

small selection of cases which specifically deal with dogwhistle tactics. The last section discusses cases considered by the Meta Oversight Board involving humour and implicit messaging.

(Dog) Whistling All the Way to Court

These case examples are drawn from various jurisdictions and different forms of communication, the common factor being that they all use dogwhistle tactics. They illustrate the various judicial and quasi-judicial approaches to implicit expression. These illustrate how high legal thresholds generally are, and the sometimes-differing processes the courts, media regulators and moderators use to assess potentially harmful humour.

The first two domestic cases from the England and Wales High Court will illustrate the difficulty of establishing thresholds for dogwhistling and decoding implicit language. Then follows a discussion of the UK's first Twitter defamation case. Next, cases brought to the European Court of Human Rights (ECtHR) will be considered alongside a British television broadcast investigated by the media regulator Ofcom. The section finishes with examples of Meta Oversight Board decisions on contentious online humorous posts. A distinction should be made between hard law as applied by States/Supranational Courts and the self-regulation of platforms through policies drawn up by private actors/companies such as Meta.

Two cases documented in the British and Irish Legal Information Institute establish how dogwhistle tactics make it difficult for the court to judge intent. Whilst not humour related – and therefore the facts of the case will not be detailed – they illustrate how the judge needs to interpret complex implicit language. The judges' arguments could equally be applied to racist or antisemitic jokes.

United Kingdom

The UK cases are *Miller & Anor v Turner* and *Erlam & Anor v Rahman & Anor.*[6] In *Erlam*, the judge noted that a completely 'innocent, indeed anodyne, statement' could be claimed to contain a coded, racist message when it did not and there was no intention for it to do so. The judge considered that in this case the coded racism existed only in the mind of the person accusing his opponent of racism [para.197]. The judgment is a reversal of the concept of the dogwhistle. It illustrates that judges are aware of over-reach in

[6] *Miller & Anor v Turner* [2023] EWHC 2799 (KB); *Erlam & Anor v Rahman & Anor* [2015] EWHC 1215 (QB). *N.b.* A search for the term 'dog whistle' and 'dogwhistle' returned two cases on 19/03/2024 (a third case referred to an actual whistle for dogs).

the decoding of messaging. Moreover, this demonstrates how dogwhistles are open to interpretation, and how useful a closer reading and consideration of the intent and implications of the speaker and implicit language can be in establishing a consistent approach to judicial reasoning (see Godioli, Young and Fiori, 2022 and Godioli et al, 2025) specifically regarding humour cases and consistency.

In the second of the UK cases, *Miller,* the judge dismissed the case but concluded that, whilst defamatory, the contested statements did not reach the threshold of serious harm to the claimant's reputation [para.163]. According to the judge, there had been 'macabre imagery and casualness around antisemitic tropes' which yielded 'a discernible signal' but which could not be objectively recognised as antisemitic hate speech or incitement to violence. She added

> That is not to say that harassment by batsqueak or dogwhistle can never amount to an actionable tort, just that I am not persuaded on the facts that there was sufficient of it to do so in this case [para.112].

Whilst this second case establishes that dogwhistles can be the cause of an actionable tort, it illustrates the high threshold. Potentially too high given the subtle nature of a dogwhistle, which is purposely multi-layered and implicit. *Miller* illustrates that these tactics can give the speaker protection from accountability (and potentially a sanction) for their words.

When a 'discernible signal' of discriminatory language is not enough for a sanction, this problem becomes greater with the addition of humour. Such signals are meant to be discernible to only a subset of the audience (potentially loud enough to incite hatred or hostility) and to unsettle the vulnerable target; the signals may not sound loud enough to a judge to reach the threshold for a legal remedy. The following sections illustrate humorous expression deemed legally contentious. These illustrate where attempts were made to pass off discriminatory or defamatory messaging as being 'just a joke'.

McAlpine v Bercow[7] is a defamation case in which the defendant used implicit ironic language. Sally Bercow was at the time a high-profile public figure and wife to the Speaker of the House of Commons. She posted this tweet, 'Why is Lord McAlpine trending? *Innocent face*.' Lord McAlpine sued for defamation. The reason for the legal action was that a serious accusation of historical child abuse carried out by a senior political figure had ignited media speculation as to the identity of the perpetrator. Lord McAlpine was a senior political figure at the time of the abuse. The readers of the tweet

[7] *McAlpine v Bercow* [2013] EWHC 1342 (QB).

needed this additional knowledge to understand the implications of the tweet. Ms Bercow denied that her tweet meant anything beyond a neutral one, wondering why Lord McAlpine was trending on Twitter, whereas the claimant believed the meaning conveyed was that he was a paedophile who had abused children. The judge looked at the question and the implication of *innocent face*. One party argued that *innocent face* was sincere, whilst the other party argued that it was used ironically. The court found that the words were defamatory, as the expression '*innocent face*' implied that the question was 'insincere and ironic'.

Ms Bercow's tweet exemplifies the difficulties of defining the meaning of a given text, because of its implicit and allusive nature, and her use of playfulness, to deny the defamatory meaning. The effect of it being online helps the deniability as the tweet can be read in isolation without the benefit of context. Those who would not have been aware of the news story would not have understood the tweet.

European Court of Human Rights

Humour is not only used to try to hide defamation but is also an effective way to reinforce stereotypes and to incite hatred (Butler, 2023: 23-24). The next case concerns an alleged 'joke' by a political activist who was convicted in France of using implicit racial discrimination by playing on the dual meaning of the French verb *voler*, which can be understood as either 'to fly' or 'to steal', *Le Pen v. France*.[8] Jean-Marie Le Pen was the honorary president of the right-wing political movement Front National. One of his speeches was posted on their website, where he had joked about Roma people who would not integrate into society but would 'fly [steal] naturally'.

An anti-racism group brought a case against Le Pen for the offence of publicly insulting a group of persons based on their ethnicity. He was found guilty and fined because of both 'intrinsic and extrinsic elements of the message'. Also, the judgment held that the statement revealed an 'offensive stereotype' and the limits of freedom of expression had been exceeded. He lodged an appeal using the argument of plausible deniability. He suggested 'the impugned remarks created doubt as to the meaning which the prosecution had given to them, which could simply mean that it is in the nature of the Roma in Eastern Europe to move' [para.9]. His appeal was dismissed.

Le Pen then turned to the ECtHR, alleging his rights to a fair trial and presumption of innocence and to freedom of expression had been breached. He argued he had not been allowed 'a few notes of humour', as it was common ground that the choice of wording was a pun. The Court concluded his comments were not limited to humour but

[8] *Le Pen v. France* (45416/16 – declared inadmissible 28th February 2017).

had rather been intended to stereotype and stigmatise the Roma community [para. 7]. Here, the Court recognised humour was used to attempt to shield the speaker from accountability, and his defence of humour, akin to 'it was just a joke', was not a valid argument to counter intentional discriminatory stereotyping. The Court concluded his case inadmissible as his words were not worthy of protection under the European Convention on Human Rights.

These types of jokes are used to try to normalise far-right movements, to make discriminatory behaviour more accessible (Schwarzenegger and Wagner, 2018). *Le Pen* and another case, *M'Bala M'Bala v. France*[9] exemplify this use of humour. In *M'Bala M'Bala*, a French comedian performed an antisemitic scene, inviting a Holocaust denier to appear on his comedy show, illustrating how dark humour can be applied in subtler ways, using under-coding (Simpson, 2003; Godioli and Young, 2023). In this case the Court reiterated that hate speech, when disguised as humour, can be as dangerous as direct speech.

The final ECtHR case is that of *Telo de Abreu* v. *Portugal*[10] in which a female politician was depicted as a pig with bare breasts and wearing stockings and suspenders in a blog post by a male opposition councillor. She brought a criminal complaint, and he was fined for aggravated defamation. He brought the case to the ECtHR on the grounds that this fine breached his right to Freedom of Expression.

The ECtHR considered that the cartoons were satirical commentary and part of an ongoing debate about local government. The Court acknowledged that the cartoons perpetuated 'regrettable' stereotyping around women in politics but agreed with the applicant and found a breach of his right to freedom of expression, stating that criminal sanctions in such cases could have a chilling effect on political satire. A concurring judgment discusses semiotic violence and how the sexualisation of women in politics creates further discrimination and violence against women. Here, the threshold was too high to be effective in restricting discriminatory expression, but unfortunately, cases such as this have a negative effect on women who participate in politics. Whilst the courts might want to protect against a chilling effect on political satire, they also need to consider the chilling effects on women who wish to involve themselves in the political sphere.

The internet has helped in focusing attention on women's experiences of violence, with the fourth wave of feminism taking an increasingly digital form, creating further opportunities for misogynistic attacks (Barker and Jurasz, 2019). These put up barriers

[9] *M'Bala M'Bala v. France* (25239/13 – declared inadmissible 20th October 2015).

[10] *Telo de Abreu v. Portugal* (42713/15, 7th June 2022).

for women who equally want to participate in public and political life but do not because of the fear of reprisal. Often, women are not challenged on their political views but rather attacked on their femininity. The Court should consider that women involved in politics are more likely than men to be the target of 'sexualized and gender-denigrating slurs', hostility, threats and violence (Håkansson, 2024).

An example of this type of misogyny, and a good illustration of how different types of regulatory oversight deliver different results, can be seen in a UK television programme. The actor and political activist Laurence Fox made sexist, misogynistic and offensive comments about the female political journalist Ava Evans. This prompted 8,867 complaints to the regulator Ofcom. Fox's comments included, 'show me a single self-respecting man that would like to climb into bed with that woman ever, and 'We don't need these sorts of "feminist 4.0". They're pathetic and embarrassing. Who'd want to shag that?'. His comments were predominantly unchallenged by the presenter, who laughed along with Fox, giving the impression that this was just jocular banter, and commented: '[...] she's [Evans] a very beautiful woman Laurence, very beautiful', as if referring to her as being physically attractive would mitigate the misogyny and offensive nature of the statements. The regulator recognised this, that it was a personal and not professional attack on a woman, and the comments were degrading, demeaning and misogynistic. Ofcom found the programme in breach of the Broadcasting Code (Ofcom, 2024).

Similar reductive depictions of women are used throughout the 'Manosphere' to represent women. Women are discredited through the use of sarcasm, sexualised imagery and humour often incorporating themes of hatred and violence (Scotto di Carlo, 2023). This can be regulated on broadcast television with stricter rules on discriminatory content, whilst it is unlikely that the expression would necessarily have reached the threshold required for a legal challenge were it delivered on an online platform. There is an opportunity for online humorous speech, which includes harmful but legal content, to be better addressed by the platforms to protect vulnerable groups in society (Matamoros-Fernández, Bartolo and Troynar, 2023). Misogynistic actors and those on the far-right have become more sophisticated at ways to recruit people to their cause (Cockerill, 2019). They use provoking, amusing and offensive content to gain attention. They advance the backlash against 'woke' to make white males feel victimised, using this to fuel white supremacy ideology. Again, they employ implicit language to evade censorship (Bhat and Klein, 2020: 152). Humour is used to 'mainstream' content which would otherwise be considered extreme (Schmid et al, 2024).

The next section considers recent decisions on humour, freedom of expression and implicit language online, illustrating the complexity of moderating hateful messages

masked with the veneer of humour. These examples are taken from Meta Oversight Board decisions, dubbed Meta's 'supreme court' (Yurieff, 2021), as judged against Meta's Community Standards. All this content had been through human moderation review, and uses communication associated with humour such as satire, caricature or cartooning.[11]

Curtains, Conspiracy, Rats and Review

As discussed in previous sections, there are different levels of regulation of expression. Television programmes need to comply with the law and editorial codes; online users' posts need to comply with both the applicable law and any additional platform policies. The positive obligations to protect freedom of expression under international law only concern State responsibility. Online platforms are private businesses and, because of this, are under no obligation to protect freedom of expression.

The reach of popular platforms like Facebook and X means that moderation practices not only inform the limits of freedom of expression online but, because of algorithms, also lead users towards content which reinforces their views (York and Zuckerman, 2019: 138), and this can be damaging to democracy (Sunstein, 2008). The UN Guiding Principles on Business and Human Rights state that business enterprises should take adequate measures to address human rights impacts to prevent, mitigate, and remedy them where appropriate (United Nations Office of the High Commissioner for Human Rights, 2012). However, this has been framed within X and Facebook as an obligation to protect users against government overreach (York and Zuckerman, 2019: 181).

The European Union's Digital Services Act, adopted in 2022, introduced obligations for larger platforms and search engines to identify, analyse and assess any systemic risks within the EU. These include spreading illegal content, any negative effects on the exercise of fundamental rights, in particular the fundamental rights to human dignity. It contains obligations for freedom of expression and information and non-discrimination. Other systemic risks include negative effects on civic discourse and gender-based violence (Article 34, Risk assessment). These systemic risks are important to address, as very large online platforms have significant power to shape both individuals and society (Schwarz, 2021). Of itself, any illegitimate censorship by either big corporations or the State could be problematic. However, it is not illegal for private corporations to ask their customers to agree to codes of conduct when using

[11] Please note these decisions were all made before Mark Zuckerburg of Meta announced the company would be over-hauling its content moderation policies on January 7th 2025. All references to the policies are to those which were in place at the time of the decisions.

their services and to caution them in some way when they breach those codes. One could draw comparisons to editorial policies, such as those used by broadcasters. Of course, the difference is that the contributors to online forums are not paid, trained professionals, but rather the general public.

The Meta Oversight Board can review and overrule moderation decisions on Facebook, Threads and Instagram. It has the principle of helping Meta 'balance free speech and safety'. The Board can make binding decisions and policy recommendations to Meta. Users can appeal to the Board to review Meta's moderation decisions (Oversight Board, about.meta.com, n.d.).

When considering context, the Oversight Board recommends using local moderators who understand geopolitical nuances and ensure 'a digital space that upholds fundamental rights, prevents harm, and encourages inclusivity and diversity' (Hatano, 2023). One could draw parallels here with the ECtHR and the margin of appreciation it gives to national courts, who may have a better understanding of the cultural and sociopolitical context and impact of certain expression. However, decoding implicit discriminatory expression gets harder as speech becomes increasingly opaque (Guercio and Caso, 2023), and if a moderator does not understand the social or geopolitical context, for example, they may miss important factors within implicit expression which would help decode the meaning. Two Oversight cases, containing implicit language and imagery in the guise of humour, are considered here.

Poland

The first case is a Polish post targeting trans people (The Oversight Board, 2024). A Facebook user had posted an image of a striped curtain in the colours of the transgender flag with a text reading 'New technology [...] Curtains that hang themselves [...] spring cleaning <3'. The post was reported a total of 12 times for Hate Speech and Suicide and Self-Injury standard (SISS). It was sent for human review against the SISS only and was found compliant. This decision was appealed and again reviewed only against the SISS and not the Hate Speech policy, and Meta upheld its decision. This was then sent as an appeal to the Board. Following the appeal, Meta removed the post on the grounds that it violated both the Hate Speech policy and SISS.

The Board considered that this post was an implicit call advocating that transgender people should commit suicide. The references to 'curtains that hang themselves' and 'spring cleaning' had not been decoded by the human reviewers as equating to a celebration of trans people committing suicide. This is especially surprising as the poster had described himself as a transphobe in his bio. The Board found the content violated the Hate Speech policy. Hate speech is prohibited

on Facebook on the grounds that it 'creates an environment of intimidation and exclusion, and in some cases may promote offline violence'. Facebook defined hate speech as a direct attack against people based on protected characteristics (Oversight Board, transparency.meta.com, n.d.).

The Board considered that the coded references to suicide alongside the transgender flag, which was a visual representation of a protected group, were 'malign creativity' used to target the LGBTQIA+ community using hateful or harassing 'posts and memes they defend as "humorous or satirical"' (The Oversight Board, 2024). Here, the Board was concerned with the post having the further effect of chilling LGBTQA+ communities' freedom of expression. The Board stated that its case analysis was informed by the international standards of freedom of expression. The board recognised that humour and satire (and some viewers of the post had found the content humorous as the 'ha-ha' reaction emojis indicated) could be used as a tool for legitimate criticism, but 'cannot be a cover for hate speech' (The Oversight Board, 2024, 8.2). The Board's recommendations included modifying Meta's internal guidance to ensure a flag which symbolises a group defined by their gender identity is recognised as such, even where no human figure is depicted. The implicit nature of the dogwhistle was only apparent to that subset that understood who the flag represented, and the subset would have contained both transphobes and trans people.

Croatia

The second example is a Croatian Facebook post of an altered version of Disney's 'The Pied Piper' cartoon with a Croatian caption, which Meta translated as 'The Player from Čavoglave and the rats from Knin'. The cartoon used many coded references to the disappearance and murders of Serbians and further dehumanised ethnic Serbs by depicting them as rats. The cartoon was posted in December 2021 on the page of a Croatian news portal known for anti-Serb attitudes. The post raised 397 complaints. It had been reviewed by an estimated forty Croatian-speaking moderators who had concluded that the video did not violate Meta's Hate Speech policy.

This is concerning given the amount of dogwhistles there were in the post. In the cartoon the village sign of Hamelin is renamed as the Croatian city of Knin, and the narrator describes how the rats take over the city by harassing and persecuting the residents. After this, a piper with a magic flute arrives from the Croatian village of Čavoglave and leads the rats from the city. What bears significance here is that the piper comes from Čavoglave, which is a coded reference to an anti-Serb song performed by a singer from the same village. As they go, the rats sing a song with lyrics which

commemorate a leader of the Serbian resistance forces in World War II. The rats are taken away by a tractor, which is a reference to a Croatian military operation which resulted in the execution and disappearance of ethnic Serbian civilians. It was clear that these dogwhistles had been decoded by those who left comments on the post, but not necessarily by the moderators.

The Board selected the case and Meta removed the post because it violated the spirit but not the letter of the Hate Speech policy. It was later concluded that the post also violated the letter of the policy. The Board decided that the case violated both the Hate Speech and the Violence and Incitement Community Standards. The implicit references glorified the violent ethnic cleansing in Knin, and the Board considered this could encourage people to feel justified in attacking Serbians.

The Board determined that moderators misinterpreted the Hate Speech policy as requiring an explicit, rather than implicit, comparison between ethnic Serbs and rats before finding a violation (The Oversight Board, 2024). The Board clarified that the Hate Speech policy prohibited attacks on protected characteristics whether the references were implicit or explicit. This case highlights the importance of context in decoding implicit messages.

The Board affirmed that Meta could remove posts which encourage violence (The Oversight Board, 2024:8.3). The Board stated that this was justified because Meta is a company and therefore its responsibilities differ from State obligations to protect human rights. This means that the Board may approve Meta's choice to remove content from its platform using a less stringent criterion than State authorities. There is often a gap in platform policies regarding the moderating of humorous expression and the interpretation of implicit messages as contrary to platform policies (Matamoros-Fernandez, 2023). Most platforms do not provide clear definitions for humour in its various forms and genres, and this adds to the interpretive challenges posed by different cases. This, in turn, feeds into problems for the human moderators—let alone automated content moderation—because of the possible multiple interpretations of humour (Aïmeur, Amri and Brassard, 2023).

Of all the large online platforms, Meta does have guidelines regarding humour, incorporating a satire exception into both Hate Speech and Dangerous Organisations and Individuals Community Standards. But more recent Oversight Board Cases show moderators still have some difficulties with interpreting satire. In non-binding policy advice, the Board recommended Meta put 'adequate procedures in place to assess satirical content and relevant context properly' to recognise when users share hateful content to condemn it and raise awareness rather than promoting a hateful ideology. Hopefully, the satire exception will improve future decisions concerning the

appropriateness of posts which use implicit speech and humour to raise awareness of important socio-political issues, and the moderators will be able to differentiate between these and dogwhistles.

Conclusion

Currently, there is no single solution which might effectively counter the challenges of socio-political dogwhistles which fall within the category of 'lawful but awful' expression. Keller (2022) discusses these issues with US internet regulation, but within the framework of the First Amendment. This makes regulating speech more difficult than it is within countries discussed here, where domestic law provides protections against explicit hate speech and expression which clash with the European Convention rights of citizens.

The aim of dogwhistles is usually to deepen societal divisions and thus damage democracy. The enormous amount of implicit discriminatory online expression makes it impossible to regulate through the courts, and so there needs to be effective regulation through online platform policy. The courts and regulators seem to have a better understanding of implicit expression than moderators. This may be because they have time to consider the nuances of humorous speech and to determine the target and intent of the speaker through a more stringent analysis. Up to now, the policies for platforms such as X and those owned by Meta have gone further than applicable law in countering hateful implicit expression, and this has been useful for the courts. More recently, X introduced Community Notes as a way of countering misinformation, which can skew information and manipulate people. As Matamoros-Fernández and Jude (2025) discuss, community notes have limits and are criticised for overlooking the potential of online harms emanating from the intersection of disinformation and humour.

Artificial intelligence moderation can go some way towards addressing online hate speech, but coded content can go undetected for long enough for it to be widely disseminated. The platforms need to improve and expand their use of human moderators who can better consider words in context. But intent is sometimes difficult to ascertain, and, like automated moderation, human moderation can be fallible.

Where a judge does not decode the underlying message or take the meaning as anything other than at face value, then it is likely the expression will not reach the legal threshold for restriction. Therefore, implicit messaging which manipulates the electorate, targets vulnerable groups, or contributes to social unrest is not necessarily captured by a legal remedy. Any remedy should combine a better understanding of the traits of dogwhistling tactics through a more effective and consistent decoding of implicit language, with robust platform policies and media literacy initiatives.

Quaranto (2022: 32) advocates for a practice-focused approach to understanding coded speech. This would entail 'focusing on the relations between and structures of multiple interactive practices, practices that are temporally extended, socio-politically shaped, historically and materially embedded'. Godioli et al (2022: 2257) discuss the importance of noting the context of the expression, given 'the subjectivity and slipperiness of humor'. This could help gauge intent and establish if humour is being used to disguise incitement to hatred or hostility, or to highlight and denounce another's discriminatory or bigoted behaviour. This does not change the law, though, as the court simply has to interpret and apply the law to any given case. Yet *Le Pen* exemplifies how an understanding of the tools of discriminatory humour can influence the outcome.

New legislation, such as the European Digital Services Act, should consider sociopolitical dogwhistles within its systemic risks because of the negative impact on society. One of the main goals of this legislation is to reduce harmful content (Turillazzi et al., 2023) and, as shown, dogwhistles are harmful and exploit emotions. As has been pointed out, much of this type of media is not illegal but is manipulative and difficult to take down through legal means because it does not reach the threshold of legal tests. Current outcries when far-right groups are deplatformed, and the so-called culture-war-based arguments that speech is only 'free' when it conforms to societal norms, mean that taking a heavy-handed approach has the potential to create a backlash (Farries, Kerrigan, and Siapera, 2024).

Given the difficulties of rapidly responding to problematic online content, the suggestion of further improving media literacy amongst the general population might be the quickest to put into action, but this is not a new idea. In 2008, the European Parliament published an overview of media literacy, defining it as being able to critically analyse media messages. It should help citizens to 'avoid or challenge media content and services that may be unsolicited, offensive or harmful' and use media effectively in 'the exercise of their democratic rights and civic responsibilities' (Policy Department Structural and Cohesion Policies, 2008). Media literacy has been taught in schools in many countries for several years.

The Council of Europe has stated that it is of utmost importance that people develop media and information literacy with objectives that include tools to empower people of all ages and backgrounds. In 2022, it launched *The Digital Era? Also my Era! Media and information literacy, a key to ensure seniors' rights to participate in the digital era*. Many young people are already equipped with some media literacy skills. Expanding these further and reaching an older audience who were not taught media literacy when they were at school is a positive step. This gives online users the tools to read, understand, recognise and (hopefully) reject any implicit discriminatory messaging. The Council of

Europe has stressed that people of all ages should be taught media literacy and critical thinking, with initiatives tailored to each nation (Council of Europe, 2024).

In cases which could be defamatory or amount to illegitimate hate speech or incitement, a better understanding of the law would potentially curb the enthusiasm of people posting illegal content. Whilst much scholarly work emphasises the harm of discriminatory expression and the way in which moderation and legislation might alleviate this, a further suggestion is that a programme of legal literacy is incorporated into media literacy initiatives and further research undertaken to build on the concept of legal consciousness around social media platforms. Legal consciousness has been defined as 'the ways in which people experience, understand, and act in relation to law' and 'some legal consciousness research demonstrates the extent to which people do not invoke or think about the law at all—or perceive it to be irrelevant' (Chua and Engel, 2019: 336).

It is not just those in editorial positions or the audience who need to be aware of a contentious post's meanings, but the post writers should be aware of the potential legal consequences of their words. Whilst some might reject the necessity of acting within the law, for others the knowledge of the legal limitations on expression might deter them from indulging in hate speech in the form of discriminatory humour.

Implicit hate speech is recognised as harmful to social cohesion and democratic debate, so there is a pressing need to counter it. In the UK, the Online Safety Act came into force in May 2025. This means that platforms had to implement measures to remove illegal content quickly once they were aware of it (Ofcom, 2025). Although this still leaves the so-called Big Tech companies as the arbiters of freedom of speech when judging implicit speech online.

Platforms owned by private companies can limit legitimate speech, and some governments might see an opportunity to implement further media regulation to stymie important socio-political speech or stifle satirical criticism. With the recent announcement from Meta that it is ceasing fact-checking and has changed some of its platform policies, further research needs to be done to compare the differences in the policies and to see what effect these might have on the user experience and the world at large.

Whilst there are a number of initiatives on a national level, there is no consistent European-wide (or indeed global) approach. Not everyone will be reached by these initiatives; indeed, it is likely that not everyone will want to be reached. There will always be bad actors seeking to exploit online users through manipulative means with little regard for the law and under the guise of 'it was only a joke'.

Competing Interests

The author has no competing interests to declare.

References

African Charter on Human and Peoples' Rights adopted 1981 [online] https://www.african-court.org/wpafc/wp-content/uploads/2020/04/AFRICAN-BANJUL-CHARTER-ON-HUMAN-AND-PEOPLES-RIGHTS.pdf [Last Accessed 08 September 2025].

Aïmeur, E, Amri, S and **Brassard, G** 2023 Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining* 13(30), pp.1–36. https://doi.org/10.1007/s13278-023-01028-5

Ajder, H and **Glick, J** 2023 Just joking! Deepfakes, satire and the politics of synthetic media. *Witness and Co-Creation Studio MIT Open Documentary Lab* https://cocreationstudio.mit.edu/wp-content/uploads/2021/12/JustJoking.pdf [Last Accessed 22 June 2025].

Åkerlund, M 2021 Dog whistling far-right code words: the case of 'culture enricher' on the Swedish web. *Information, Communication & Society*, 25(12), pp. 1808–1825. https://doi.org/10.1080/1369118X.2021.1889639

American Convention on Human Rights 1969 [online] https://www.oas.org/en/iachr/mandate/basics/3.american%20convention.pdf [Last Accessed 08 September 2025].

Baider, F 2022 Covert Hate Speech, Conspiracy Theory and Anti-semitism: Linguistic Analysis Versus Legal Judgement. *International Journal for the Semiotics of Law – Revue internationale de Sémiotique juridique*, 35, pp. 2347–2371. https://doi.org/10.1007/s11196-022-09882-w

Baker, T and **Scott, J** 2024 Anderson defects to Reform UK – as Beth Rigby challenges three-word answer to question. *Sky News* https://news.sky.com/story/lee-anderson-former-conservative-deputy-chair-defects-to-reform-uk-13092518 [Last Accessed 26 June 2024].

Bandrowski, A C 2024 Words Not Said: Can the Brandenburg Incitement Test Cope with Coded Speech? *Boston College Law Review*, 65, pp. 1483–1520. https://bclawreview.bc.edu/articles/3136

Barker, K and **Jurasz, O** 2019 Online Misogyny: A Challenge for Digital Feminism? *Journal of International Affairs*, 72(2), pp. 95–114. https://www.jstor.org/stable/26760834

Bayer, J 2021 High-impact hate speech by persons of authority: A lower threshold needed? *Hungarian Journal of Legal Studies*, 61(3), pp. 269–284. https://doi.org/10.1556/2052.2020.00003

Beck, D and **Spencer, A** 2025 (Un)Funny Against All Odds: The Changing Landscape of Humour in Politics. *Alternatives*, 50(1), pp. 3–17. https://doi.org/10.1177/03043754241290911

Bhat, P and **Klein, O** 2020 Covert Hate Speech: White Nationalists and Dog Whistle Communication on Twitter. In Bouvier, G and Rosenbaum, JE (eds) *Twitter, the Public Sphere, and the Chaos of Online Deliberation*, Palgrave Macmillan, Cham. pp. 151–172. https://doi.org/10.1007/978-3-030-41421-4_7

Billig, M 1995 *Banal Nationalism*. London: Sage.

Boholm, M and **Sayeed**, A 2023 *Political dogwhistles and community divergence in semantic change*. Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change *Association for Computational Linguistics*.

Butler, N 2023 The *Trouble with Jokes: Humour and Offensiveness in Contemporary Culture and Politics*. Bristol: Bristol University Press.

Chua, L J and **Engel, D M** 2019 Legal Consciousness Reconsidered *Annual Review of Law and Social Science* 15, pp.335–353. https://doi.org/10.1146/ANNUREV-LAWSOCSCI-101518

Cockerill, M 2019 Convergence on Common Ground: MRAs, Memes and Transcultural Contexts of Digital Misogyny. In Ging, D and Siapera, E (eds) *Gender Hate Online*, Palgrave Macmillan, Cham. pp. 87–110.

Council of Europe 2022 Also my Era! Media and information literacy, a key to ensure seniors' rights to participate in the digital era. https://rm.coe.int/digital-literacy-for-seniors-print/1680a6ce9e [Last Accessed 28/04/2025].

Council of Europe 2024 *Media literacy and the Empowerment of Users*. [online] European Audiovisual Observatory. Available at: https://www.obs.coe.int/en/web/observatoire/-/new-report-media-literacy-and-the-empowerment-of-users [Last Accessed 15 Jan. 2025].

Elder, C-H 2024 *Pragmatic Inference*. (Cambridge Elements in Pragmatics) Cambridge: Cambridge University Press.

European Parliament 2008 An Overview of Media Literacy: *Policy Department Structural and Cohesion Policies*. https://www.europarl.europa.eu/RegData/etudes/note/join/2008/397254/IPOL-CULT_NT(2008)397254_EN.pdf [Last Accessed 15 Jan. 2025].

Farries, E, Kerrigan, P and **Siapera, E** 2024 The Platformisation of Cancel Culture. *Television & New Media*. https://doi.org/10.1177/15274764241277469

Feinberg, A 2017 This Is The Daily Stormer's Playbook. Huffpost 13/12/2017 [online] https://www.huffingtonpost.co.uk/entry/daily-stormer-nazi-style-guide_n_5a2ece19e4b0ce3b344492f2 [Last Accessed 27/01/2025].

Fielitz, M and **Reem, A** 2021 It's Not Funny Anymore. Far-Right Extremists' use of Humour. *Publications Office of the European Union*. <https://home-affairs.ec.europa.eu/system/files/2021-03/ran_ad-hoc_pap_fre_humor_20210215_en.pdf> [Last Accessed 26/06/2024].

Ganesh, B 2018 The Ungovernability of Digital Hate Culture. *Journal of International Affairs* 71(2), pp. 30–49. https://www.jstor.org/stable/26552328

Gillespie, T 2020 Content moderation, AI, and the question of scale. *Big Data & Society*, 7(2), pp.1–5 https://doi.org/10.1177/2053951720943234

Godioli, A, Jacques, S, Young, J and **Matamoros Fernandez, A** 2025 What's in a Joke? Assessing Humor in Free Speech Jurisprudence. *Forum for Humor and the Law/Columbia Global Freedom of Expression*. https://doi.org/10.5281/zenodo.15383543

Godioli, A and **Young, J** 2023 *Humor and Free Speech: A Comparative Analysis of Global Case Law*. Columbia Global Freedom of Expression, Special Collection. https://doi.org/10.5281/zenodo.8105760

Godioli, A, Young, J and **Fiori, B M** 2022 Laughing Matters: Humor, Free Speech and Hate Speech at the European Court of Human Rights. *International Journal for the Semiotics of Law – Revue internationale de Sémiotique juridique*, 35(6), pp. 2241–2265. https://doi.org/10.1007/s11196-022-09949-8

Guercio, N L and **Caso, R** 2023 Special Issue on Dogwhistles. *Manuscrito*, 46(3). https://doi.org/10.1590/0100-6045.2023.v46n3.nr

Håkansson, S 2024 Explaining Citizen Hostility against Women Political Leaders: A Survey Experiment in the United States and Sweden. *Politics & Gender* 20(1): pp. 1–28.

Hatano, A 2023 Regulating Online Hate Speech through the Prism of Human Rights Law: The Potential of Localised Content Moderation. *The Australian Year Book of International Law Online*, 41(1), pp. 127–156. https://doi.org/10.1163/26660229-04101017

Henderson, R and **McCready, E** 2018 How Dogwhistles Work. *New Frontiers in Artificial Intelligence*, pp. 231–240. https://doi.org/10.1007/978-3-319-93794-6_16

Holmes, J and **Marra, M** 2002 Over the Edge? Subversive Humor Between Colleagues and Friends. *Humor: International Journal of Humor Research*, 15(1), pp. 65–87.

Hutchinson, A September 25 2024 *Data Shows X Is Suspending Far Fewer Users for Hate Speech*. [online] Social Media Today https://www.socialmediatoday.com/news/data-shows-x-suspending-far-fewer-users-hate-speech/728136/ [Last Accessed 27/01/2025].

Jacques, S 2019 *The Parody Exception in Copyright Law.* Oxford: Oxford University Press.

Kantrowitz, A October 2, 2016 *Racist Social Media Users Have a New Code to Avoid Censorship*. [online] BuzzFeed News. https://www.buzzfeednews.com/article/alexkantrowitz/racist-social-media-users-have-a-new-code-to-avoid-censorshi [Last Accessed 14 Jan. 2025].

Kasimov, A Johnston, R and **Heer, T** 2023 'Pepe the frog, the greedy merchant and #stopthesteal': A comparative study of discursive and memetic communication on Twitter and 4chan/pol during the insurrection on the US Capitol. *New media & Society:* pp. 1–24. https://doi.org/10.1177/14614448231172963

Keller, D 2022 Lawful but Awful? Control over Legal Speech by Platforms, Governments, and Internet Users. *University of Chicago Law Review Online*, pp. 1–12 Lawful but Awful? Control over Legal Speech by Platforms, Governments, and Internet Users | The University of Chicago Law Review [Last Accessed 15/08/2025].

Kotecha, S 10 August 2024 *Riots resurface memories of racist violence for British Asians – with glimmer of hope*. [online] BBC.com. https://www.bbc.com/news/articles/ckg2r3lxzedo [Last Accessed 27th Jan 2025].

Lewis, B and **Marwick, A E** 2017 *Media Manipulation and Disinformation Online.* Data & Society [online] https://datasociety.net/library/media-manipulation-and-disinfo-online/ [Last Accessed 27th Jan 2025].

Marwick, A E 2021 Morally Motivated Networked Harassment as Normative Reinforcement. *Social Media + Society*, 7(2), pp.1–13 https://doi.org/10.1177/20563051211021378

Matamoros-Fernández, A 2023 Taking Humor Seriously on TikTok. *Social Media + Society*, *9*(1). https://doi.org/10.1177/20563051231157609 [Last Accessed 01/09/2025].

Matamoros-Fernández, A, Bartolo, L and **Troynar, L** 2023 Humour as an online safety issue: Exploring solutions to help platforms better address this form of expression. *Internet Policy Review*, 12(1). https://doi.org/10.14763/2023.1.1677

Matamoros-Fernández, A and **Jude, N** 2025 The importance of centering harm in data infrastructures for 'soft moderation': X's Community Notes as a case study. *New Media and Society*, 27(4), pp.1987–2011. https://doi.org/10.1177/14614448251314399

McKevitt, S 2019 Persuasive politics: why emotional beats rational for connecting with voters. [online] *The Conversation*. Available at: https://theconversation.com/persuasive-politics-why-emotional-beats-rational-for-connecting-with-voters-116098 [Last Accessed 27/01/2025].

Mendel, T 2010 Hate Speech Rules Under International Law. *Centre for Law and Democracy*. https://www.law-democracy.org/wp-content/uploads/2010/07/10.02.hate-speech.Macedonia-book.pdf. [Last Accessed 27 January 2025].

Mistry, P 2024 BBC News, 'Keyboard warrior' jailed for part in UK disorder. https://www.bbc.co.uk/news/articles/c5y3gre3y9yo [Last Accessed 5 May 2025].

Ofcom 2024 Misogynistic comments on Dan Wootton Tonight broke offence rules. https://www.ofcom.org.uk/tv-radio-and-on-demand/broadcast-standards/misogynistic-comments-on-dan-wootton-tonight-broke-offence-rules [Last Accessed 5 May 2025].

Ofcom 2025 Statement: Protecting People from Illegal Harms Online. https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/statement-protecting-people-from-illegal-harms-online [Last Accessed 24 April 2025].

Oversight Board | **Meta** *Our Work* https://about.meta.com/actions/oversight-board-facts/ [Last Accessed 27 January 2025]

Oversight Board | **Meta** 2022 *Hateful Conduct* | *Transparency Center.* [online] https://transparency.meta.com/en-gb/policies/community-standards/hate-speech [Last Accessed 15th January 2025].

Oversight Board | **Meta** 2024 *Knin cartoon* | *Oversight Board*. [online] https://www.oversightboard.com/decision/fb-jrq1xp2m/ [Last Accessed 15 January 2025].

Oversight Board | **Meta** 2024 *Post in Polish Targeting Trans People* | *Oversight Board*. [online] https://www.oversightboard.com/decision/fb-uk2rus24/ [Last Accessed 15 January 2025].

Pérez, R 2017 Racism without Hatred? Racist Humor and the Myth of 'Colorblindness'. *Sociological Perspectives* 60(5), pp.956–974. https://doi.org/10.1177/0731121417719699

Quaranto, A 2022 Dog whistles, covertly coded speech, and the practices that enable them. *Sythese* (200:330) https://doi.org/10.1007/s11229-022-03791-y

Rowbottom, J 2012 To Rant, Vent and Converse: Protecting Low Level Digital Speech *The Cambridge Law Journal* 71(2), pp. 355–383. https://doi.org/10.1017/s0008197312000529.

Saul, J M 2024 *Dogwhistles and Figleaves*. Oxford: Oxford University Press.

Sayeed A, Breitholtz E, Cooper R, Lindgren E, Rettenegger, G, and **Rönnerstrand B** 2024 The utility of (political) dogwhistles: A life cycle perspective. *Journal of Language and Politics*. https://doi.org/10.1075/jlp.23047.say

Schmid, U K, Schulze, H, and **Drexel, A** 2024 Memes, humor, and the far right's strategic mainstreaming. *Information, Communication & Society*, 28(4): pp. 537–556. https://doi.org/10.1080/1369118X.2024.2329610

Schwarz, O 2021 *Sociological Theory for Digital Society*. John Wiley & Sons.

Schwarzenegger, C and **Wagner, A** 2018 Can it be hate if it is fun? Discursive ensembles of hatred and laughter in extreme right satire on Facebook. *Studies in Communication | Media*, 7(4), pp.473–498. https://doi.org/10.5771/2192-4007-2018-4-473

Scotto di Carlo, G 2023 An analysis of self-other representations in the incelosphere: Between online misogyny and self-contempt *Discourse & Society*, 34(1): pp. 3–21.

Simpson, P 2003. *On the Discourse of Satire: Towards a Stylistic Model of Satirical Humour*. Amsterdam: John Benjamins.

Simpson, P 2023 Irony and Its Consequences in the Public Sphere. In: Gibbs, Jr RW, Colston HL (eds.) *The Cambridge Handbook of Irony and Thought*. Cambridge Handbooks in Psychology. Cambridge: Cambridge University Press. pp. 112–128.

Strong, C, Owen, K and **Mansfield, J** 2023 *Understanding experiences of minority beliefs on online communication platforms*. Office of Communications [online] https://www.ofcom.org.uk/__data/assets/pdf_file/0019/268102/understanding-experiences-minority-beliefs.pdf [Last Accessed 14 January 2025].

Sunstein, C 2008 Democracy and the Internet. In van den Hoven, J & Weckert J (eds.) *Information Technology and Moral Philosophy*. Cambridge: Cambridge University Press. pp. 93–110.

Suzor, N P 2019 *Lawless: The Secret Rules That Govern Our Digital Lives*. Cambridge: Cambridge University Press.

Tsakona, V 2024 *Exploring the Sociopragmatics of Online Humor (Topics in Humor Research, 12)*. John Benjamins Publishing Company.

Turillazzi, A, **Taddeo, M, Floridi, L** and **Casolari, F** 2023 The digital services act: an analysis of its ethical, legal, and social implications. *Law, Innovation and Technology*, 15(1): pp. 83–106.

United Nations Office of the High Commissioner for Human Rights (n.d.). *OHCHR | OHCHR and freedom of expression vs incitement to hatred: the Rabat Plan of Action*. [online] https://www.ohchr.org/en/freedom-of-expression [Last Accessed 14 January 2025].

United Nations Office of the High Commissioner for Human Rights 1966 International Covenant on Civil and Political Rights General Assembly resolution 2200A(XXI). https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights [Last Accessed 26 June 2024].

United Nations Office of the High Commissioner for Human Rights 2012 *Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework*. [online] https://www.ohchr.org/en/publications/reference-publications/guiding-principles-business-and-human-rights [Last Accessed 26 June 2024].

United Nations Office of the High Commissioner for Human Rights 2020 One-pager on 'incitement to hatred' https://www.ohchr.org/en/documents/tools-and-resources/one-pager-incitement-hatred-rabat-threshold-test [Last Accessed 26 June 2024].

X's policy on hateful conduct | X Help [Last Accessed 1 October 2024].

York, J C and **Zuckerman, E** 2019 Moderating the Public Sphere. In Jørgensen, RF (ed) *Human Rights in the Age of Platforms*. Cambridge Massachusettes: The MIT Press.

Yurieff, K 2021 *Facebook's 'supreme court' just ruled against Facebook*. [online] CNN. https://edition.cnn.com/2021/01/28/tech/facebook-oversight-board-first-decisions/index.html [Last Accessed 14 January 2025].

Zinigrad, R 2024 Laughing Matters in Courts: Humor's Role in Normalizing Hate Speech. *Alternatives*, 50(1): pp. 33–51. https://doi.org/10.1177/03043754241244891