Jibe and Prejudice: On Satire, Deception, and Harm

Alberto Godioli & Luisa Fernanda Isaza-Ibarra (University of Groningen / ForHum)

ForHum *(Forum for Humor and the Law) is an international platform promoting a fair approach to humorous expression in free speech adjudication, in light of insights from legal scholarship as well as humour research from the humanities and social sciences. This public comment discusses the Oversight Board case AI video faking UK politician's immigration views.*

In November 2025, a Facebook user shared an album containing several pieces of content, including an apparently **satirical video** depicting a UK Labour Party councillor who represented a constituency in Scotland. The video seems to have been generated using AI, and shows the councillor stating: "Refugees are welcome here, even if they rape our women, because white people do that too". The album also featured other videos and images targeting the councilor's political views, and anti-far-right movements at large. The post generated fewer than 50 reactions and comments, and was shared fewer than 50 times. As implied by the Oversight Board in its summary, the case raises two interconnected but distinct issues, namely:

1) The risks of **deception** and **misinformation** potentially stemming from satirical expression (especially in the form of AI-generated videos). In other words: *to what extent can the video be interpreted as a genuine statement by the UK Labour Party councillor*? From this standpoint, the target of the disputed expression is the councillor herself, while the most relevant Meta policies are those on Misinformation on the one hand, and Bullying and harassment on the other.

2) The degree to which certain forms of humour or satire expression can reinforce **discriminatory** messages (i.e., in this case, the idea that refugees "rape our women"). In this respect, the target of the video are refugees at large, with Hateful conduct being the relevant policy.

Our comment will address both issues. Based on scholarship regarding humour and satire as well as free speech and content moderation, we will tackle the complex relationship between satire and misinformation (Part 1), and between disparaging humour and harm (Part 2).

1. Satire and Misinformation

1.1. Satire, confusion and incongruity

As argued in our previous public comment to the Oversight Board, the risk of **confusion** between humorous exaggeration and reality is not an accidental side effect of satire, but one of its defining features. Satirical representation needs to be anchored in reality and bear some resemblance with the target's actual features or habits, in order for the critical message to actually come across. Linguist Paul Simpson calls the resemblance the 'echoic' element of satire, while the critical message is the 'oppositional' element.

What distinguishes satire from deception or misinformation is that this state of confusion is supposed to be **temporary**. In essence, humour and satire are a form of 'overt untruthfulness' (Dynel 2018) – meaning that they typically contain signals giving away their non-factual nature. These signals usually amount to what humour scholars call **incongruity** – that is, some kind of implausibility, achieved by means of

exaggeration, absurd juxtaposition, or distortion of commonsense logic. Let us take Jonathan Swift's *A Modest Proposal* (1729) as an example:

> I shall now therefore humbly propose my own thoughts, which I hope will not be liable to the least objection. I have been assured by a very knowing American of my acquaintance in London, that a young healthy child well nursed, is, at a year old, a most delicious nourishing and wholesome food, whether stewed, roasted, baked, or boiled; and I make no doubt that it will equally serve in a fricasee, or a ragoust.

At first, the reader might be tricked into thinking they are reading a 'serious' text, as Swift carefully imitates the style of 18th-century pamphleteers and their utilitarian logic. However, the enormity of the idea put forward by the pamphleteer (eating the children of the poorest families to solve poverty in Ireland) gives away Swift's satirical intent – i.e., to ridicule the very logic he is imitating ('echoic' phase), by exaggerating it to its absurd consequences ('oppositional' phase).

This mechanism is brilliantly described by *The Onion*'s celebrated amicus brief in *Novak v. Parma* (2022). In particular, *The Onion* stresses how satire and parody (the two terms are used interchangeably in the brief) often function by initially "**tricking people**" into believing what they are reading, only to reveal their message through incongruity:

> The phrase 'you are dumb' captures the very heart of parody: tricking readers into believing that they're seeing a serious rendering of some specific form—a pop song lyric, a newspaper article, a police beat—and then allowing them to laugh at their own gullibility when they realize that they've fallen victim to one of the oldest tricks in the history of rhetoric (pp. 8-10).

Importantly, this rhetorical trick is not only organic to satirical expression, but also serves a broader **social function**. By facing readers with their own gullibility, satire encourages us to engage more critically with information, rather than passively relying on external cues to determine what should be believed. In this way, satire can contribute to media literacy by cultivating habits of scrutiny, interpretation, and scepticism.

Moving closer to the case at hand, **satirical deepfakes** are clearly based on the same mechanism. The main difference from other forms of satire – such as a cartoon or a comedian's impression – is that the 'echoic' phase (imitating the target) is particularly realistic and potentially confusing, given the unprecedented emulative possibilities offered by AI. Nevertheless, deepfakes can be a legitimate form of political satire, as recently stressed by the U.S. District Court for the Eastern District of California in *Kohls v. Bonta* (currently pending on appeal). In this ruling, the court struck down key provisions of a California law that would have restricted certain forms of AI-generated parody and satire. The court recognised that satire often depends on its capacity to resemble authentic communication, and that the plaintiffs' "exaggerated and hyperbolic" videos formed part of a "long-held American tradition of ridiculing and criticizing candidates and elected officials across the political spectrum".

Given the higher confusion potential, it is all the more important for satirical deepfakes to display a degree of **incongruity**, setting them apart from outright deceptive content. As discussed in our previous comment (based on relevant case law), the incongruity should at least be noticeable by a "reasonable" member of the audience. Building on recent work on the human detection of deepfakes (Groh et al. 2024), we can distinguish two ways in which deepfakes might give away their fictional nature – one is *what* is being said in the video (i.e., the content), and the other is *how* the video is made (i.e., formal audio-visual cues). This allows us to identify two basic types of satirical incongruity in deepfakes:

1) *Content-based incongruity*: the deepfake is satirical because the statements or actions attributed to the people in the video are unbelievable, outlandish, or out of character. Two cases in point, taken from the WITNESS report on deepfakes and satire, are the video showing far-right French politician Marine Le Pen wearing a hijab and speaking Arabic; and the one in which former Brazilian president Jair Bolsonaro (in stark contrast to his real-life counterpart) sings about the urgency of careful hand washing during the Covid pandemic.

2) *Form-based incongruity*: the deepfake is satirical because of its formal (audio/video) features. For instance, *South Park*'s famous Donald Trump deepfake signals its satirical intent through the unrealistic imitation of the President's voice, and through the fact that the video – at least in its original context – is embedded within an animated cartoon.

How, then, does this apply to the **UK video** currently examined by the Oversight Board? Without access to the original content, it is impossible to assess the degree of form-based incongruity. The content, however, is definitely incongruous: there is no situation in which a progressive politician could realistically utter words such as "refugees are welcome here, even if they rape our women". The incongruity, in this case, is based on a millenia-old satirical mechanism, namely the *reductio ad absurdum* – mimicking your opponent's stance ("refugees are welcome") and pushing it to unacceptable extremes (condoning rape). Despite the lamentable, discriminatory nature of the message being conveyed (to be further discussed in Part 2), there is little doubt about the satirical nature of the video.

1.2. Deepfakes, political participation, and media literacy

This does not mean that the confusion caused by satirical deepfakes (even if temporary) cannot be **harmful** to the people they portray. In this regard, it is worth stressing that the councillor portrayed in the video is a woman, and that **women's participation in politics** is hindered by a disproportionately high degree of identity-based online intimidation (Jarman 2026, Gehrke 2026, Lühiste et al. 2025). Recent initiatives like the Forced to Quit project document how different forms of disinformation – including deepfakes – effectively deter women from active political engagement, thereby becoming a serious obstacle to freedom of expression and democratic participation.

In light of our discussion above, and based on the information at hand, the disputed video does not seem to qualify as identity-based harassment or disinformation *per se*. At the same time, we do acknowledge how, given its triggering topic, the video could expose the depicted councillor to online abuse by other users – which, in that event, should be addressed according to Meta's policies concerning Bullying and harassment. That being said, it is important to stress that **online harm** is a multi-layered phenomenon; rather than in binary terms (*harmful* vs *non-harmful*, *remove* vs *restore*, etc.), it should be conceived as a nuanced spectrum, calling for a wide range of proportionate countermeasures (Bartolo & Matamoros-Fernández 2023). In the case of satirical deepfakes such as the one under discussion – non-sexualized, not amounting to outright bullying or harassment, and marked by noticeable incongruity –, we believe that such countermeasures should primarily focus on counterspeech and longer-term **awareness-raising** endeavours.

Once again, a parallel with **eighteenth-century satire** might serve as inspiration. As mentioned in our previous comment, Abigail Williams' book *Reading It Wrong* (2023) establishes a persuasive parallel between the golden age of British satire and present-day online culture – both being marked by the rise of new audiences, a dramatic increase in the circulation of texts, and the proliferation of allusive satirical

expression, often playing with the risk of confusion or misunderstanding. In the framework of this comparison, it seems particularly fruitful to consider the role played by periodicals such as Joseph Addison and Richard Steele's *The Spectator* (1711-1714). As shown by Williams, *The Spectator* served an important pedagogical function in helping its public learn how to read satire – i.e. how to detect and respond to irony and ambiguity, while engaging in a broader **culture of interpretation** based on critical and conversational reading.

Applying this principle to the present, we believe it is crucial for **online platforms** to proactively foster such a culture of interpretation – including by means of prebunking (training people before they encounter misinformation, rather than correcting them afterward) and supporting media literacy projects to help readers navigate the grey areas between satire and misinformation. Within this context, synthetic media could prove to be not only a threat, but also a valuable educational tool, as in Jordan Peele's farsighted satirical deepfake on the importance of relying on "trusted news sources".

2. Humour and hate speech

In satirising the councillor's views on migration, the disputed video also advances a common anti-migration trope which, many would argue, is hateful. We will begin by highlighting that, since it criticises the stance of a public official on a matter of high public interest, the video could be considered a type of **specially protected speech** under international freedom of expression standards. Therefore, its restriction would require strong justifications. This principle is somewhat reflected in Meta's Hateful conduct policy, which states that while the company protects migrants, it also allows "criticism of immigration policies".

Regarding the **public interest** angle, a relevant precedent is the Oversight Board case on U.S, posts discussing abortion, where the Board overturned Meta's original decisions to remove three posts discussing abortion from different pro/against stances. While the posts contained rhetorical uses of violent language, the Board highlighted that discussions about abortion are a key political issue in the United States, that Facebook and Instagram are important sites for political discussion, and that "Meta has a responsibility to respect the freedom of expression of its users on controversial political issues".

That said, it is also important to recognise that humour can indeed be **hateful**, and that the "just joking" defence should not guarantee blanket immunity from consequences. Sociologists Blee and Simi (2020), who have worked on far-right and white supremacist movements, have explained that "the tactic of joking can be used as a form of double-speak to deny culpability". Building on their work, Saul (2023) also referred to these formulations as a "figleave" – like plants covering the nudity of Adam and Eve in paintings and sculpture, "just joking" arguments tenuously disguise utterances that are in fact racist. Within humour scholarship, this type of content falls under what is known as **disparaging humour**, i.e. humour that denigrates, belittles, or maligns an individual or social group (Ford & Ferguson 2004). This issue is increasingly receiving attention from scholars, variously showing how jokes can reinforce systems of oppression and discrimination (Horisk 2024, Pérez 2022).

Adding some nuance to this picture, empirical research on the social effects of disparaging humour in physical settings (Ford et al. 2015) suggests that while exposure to derogatory jokes does not necessarily *spread* prejudice, it can facilitate the expression of **existing prejudice** without fears of reprisal. In other words, being exposed to racist or sexist jibes does not automatically make people more racist, sexist, etc.; rather, people who already hold those beliefs seem to be more comfortable sharing their views. A

similar point is made by Schmid (2023) who, moving to the digital world and focusing on hateful memes, explains that: "only those with preexisting prejudices against marginalized groups found hateful memes toward them amusing, similar as research has shown that disparagement humor can reinforce only preexisting prejudices".

Turning to the contested video, how can we assess its potential to harm? To begin with, while the relevant framework in this case is the Hateful conduct policy, it is still relevant to highlight that (based on the available information) the case falls short of the **incitement** threshold set forth in the United Nations' Rabat Plan of Action. In particular:

1) The user who posted the video does not seem to hold any relevant level of social **authority** or influence;
2) At 50 reactions, comments, and shares, the post's extent of **dissemination** at the moment of the call seems minimal;
3) The post appears mostly aimed at discrediting one individual politician, rather than directly calling on an audience to act against refugees as a group. Therefore, it seems unlikely that the video will create an **imminent risk** of "actual action" against refugees.

Nevertheless, the video does advance a derogatory trope, and – despite its limited reach – can therefore act as a "**releaser**" of prejudice (Ford & Ferguson). In such cases, Meta can resort to a broad array of options, including (but not limited to) removal, contextual labels, downranking, or simply leaving the content untouched. Rather than advocating one specific approach, we wish to emphasise three general principles that may be relevant in this case, drawing on scholarship concerning freedom of expression and content moderation.

1) As already mentioned, online harm is best conceived of as a continuum rather than a binary category. Accordingly, it should be addressed through "a range of different and **proportionate** remedies" (Bartolo & Matamoros-Fernández 2023), aiming to protect users while avoiding the risk of regulatory overreach.
2) Overreliance on removal can result in a regulatory **slippery slope** (Future of Free Speech 2023), facilitating self-victimisation on the part of bad actors (Jacobs 2021, Rogers 2020, Strossen 2018), or even silencing the very groups that such policies seek to protect (Dias Oliva et al. 2021).
3) Lastly, awareness-raising and **counterspeech** – especially speech inviting users to adopt the perspective of targeted groups – can prove more effective than merely restrictive methods (Gennaro et al. 2025, Ullmann & Tomalin 2024).

The third point in particular seems to be further supported (albeit indirectly) by the empirical humour research discussed above, highlighting how disparaging humour primarily serves as a releaser of pre-existing **prejudice**, as opposed to 'infecting' those who do not already harbour discriminatory views. If this is indeed the case, restricting the disputed content may still be a valid option in some cases, but intervening on the underlying prejudice may ultimately be more effective than taking down the joke.

Funding note: *This comment reflects research carried out by the authors in the framework of the projects Humour in Court (VI.Vidi.201.111, NWO Talent Programme Vidi SSH 2020) and DELIAH: Democratic Literacy and Humour (Horizon Europe, Project n. 101177739).*