PUBLIC COMMENTS ON CASE 2023-029-FB-UA ALTERED VIDEO OF PRESIDENT BIDEN

Tech Global Institute (https://techglobalinstitute.com) is a policy lab with a mission to reduce equity and accountability gaps between technology platforms and the Global Majority. In this submission, we respond to the Oversight Board's request for public comments on the Altered Video of President Biden with specific reference to the following issues. We are making this submission because the Oversight Board's decision will profoundly impact Meta's future review of altered videos featuring political leaders, not only in the United States but also in the Global South.

Research into online trends of using altered or manipulated video content to influence the perception of political figures, especially in the United States.

Online trends around using altered or manipulated video content to influence the perception of political figures, especially in the United States, include:

- **Deepfakes and AI-driven manipulation:** Advances in artificial intelligence have made it possible to create hyper-realistic but entirely fake content. Deepfake videos use machine learning algorithms to generate fabricated videos of real people, saying or doing things they never actually said or did. Examples include:

    - A deepfake video of former US President Barack Obama calling Donald Trump "a total and complete dipshit" was shared online in 2018. Although the video was quickly debunked, it highlighted the potential for deepfakes to be used to interfere with elections.
    - Several altered images of former US President Donald Trump surfaced online showing him hugging and kissing scientist Dr Anthony Fauci. Other images show him in altercation with policemen, posing for a mugshot and in orange prison overalls, and leading a rally.
    - Several altered videos of former US President George Bush appeared online. One shows him explaining future of artificial intelligence, while another features him in a scene from *Harold & Kumar*.
    - Between 2019 and 2023, several deepfake videos of British politicians emerged online. One deepfake video shows UK PM Boris Johnson and opposition leader Jeremy Corbyn endorsing each other for prime minister, shared online in an attempt to demonstrate the potential of altered media to undermine democracy.
    - An audio recording surfaced online in October 2023 in which opposition leader Sir Keir Starmer was heard berating party staffers in a profanity-laden tirade on the first day of Labour Party conference, while in another recording he is heard saying he "hated" the city of Liverpool where the conference was held.
    - An altered image of UK PM Rishi Sunak appeared online showing him pulling a sub-standard pint at the Great British beer festival while a woman looks on with a derisive expression, while the original photo shows Sunak appearing to have pulled a pub-level pint while the person behind him has a neutral expression.
    - Altered media related to Russo-Ukraine war has circulated on social media in 2022. A deepfake video of Ukrainian President Volodymyr Zelenskyy surrendering to Russia circulated online. While the video was intended to demoralize Ukrainian troops and civilians, it was quickly debunked. Similarly, a deepfake video of Russian President Vladimir Putin appeared online claiming that Russia has won the war and that Ukraine has recognized Crimea as Russian territory. Another video from 2023 shows Putin announcing that Russia was under attack and declared martial law with a full-scale mobilization plan.
    - Other examples include: Jim Acosta, Jennifer Lawrence and Steve Buscemi, David Beckham Anti-Malaria PSA, World Leaders Sing "Imagine", Dali Museum, Bill Hader impressions, Mark Zuckerberg, Joe Rogan, Nixon and a moon landing, Queen's Christmas speech, Tom Cruise TikToks, Pennsylvania cheerleader case, and an Anthony Bourdain documentary.

    A report prepared by the Department of Homeland Security demonstrates the increasing threats of deepfake identities.

- **Shallowfakes or Cheapfakes:** This term refers to videos that have been manipulated using more basic methods compared to deepfakes. For example, these methods might involve editing out context or manipulation of video speed to create the illusion of someone slurring their words. An illustrative example is the case of altered media depicting Nancy Pelosi as appearing intoxicated and slurring her words, later revealed to be a fake but still caused significant damage to Pelosi's reputation.

- **Misleading edits:** Some video manipulations are simple edits made to take statements out of context, clip segments that can be misconstrued, or combine unrelated pieces of footage to create a misleading narrative. This can happen by:

  - **Splicing video clips together to create a false narrative**: This is a common tactic used in political campaigns and propaganda videos. For example, a video might be edited to make it appear as if a politician said something they never actually said.

  - **Adding or removing audio from a video**: This can be used to make it appear as if someone is saying something they never actually said, or to make them sound more or less enthusiastic about something.

  - **Changing the speed or pitch of audio**: This can be used to make someone sound more or less intelligent, or to make them sound more or less emotional.

  - **Manipulating images with editing software**: This can be used to make someone look older, younger, healthier, or sicker than they actually are. It can also be used to change the background of an image or to add or remove objects from an image.

- **Memes and satire:** While not always malicious, comedic or satirical videos that distort reality can sometimes be shared out of context, leading viewers to misconstrue the intent or believe in the content's veracity.

- **Political usage:** While many politicians and their supporters condemn the use of manipulated videos, there have been instances where altered videos were shared either knowingly or unknowingly by political figures or their affiliates, leading to controversy. For example, Rudolph Giuliani accidentally shared a deepfake video of Nancy Pelosi on Twitter.

While we provide examples predominantly from the United States, manipulated media, including deepfake, have become increasingly prevalent in influencing perceptions about political figures in other parts of the world. Some examples include an Indian politician using deepfake technology to win new voters, an alleged audio deepfake used to frame presidential candidate and other senior leaders in the 2023 Nigeria national election, an alleged deepfake image by former Pakistani PM Imran Khan to mislead citizens about law enforcers targeting his female supporters.

The suitability of Meta's misinformation policies, including on manipulated media, to respond to present and future challenges in this area, particularly in the context of elections.

While admitting that its misinformation policy cannot, and does not, articulate a comprehensive list of prohibited content due to the ever-evolving nature of the definition of "misinformation", Meta categorizes certain types of content that it treats as misinformation. This includes "content that is likely to directly contribute to interference with the functioning of political processes and certain highly deceptive manipulated media."

- While the policy states that it aims "to promote election […] integrity, [and] remove misinformation that is likely to directly contribute to a risk of interference with people's ability to participate in those processes," it primarily concentrates on addressing straightforward logistical misinformation, such as voting dates, locations, and eligibility, as well as participation in the census. However, the policy falls short in addressing more nuanced and insidious threats. For instance, it does not cover the sophisticated manipulation of audio and video media – such as deepfakes and synthetic media – which can fabricate speeches and actions by candidates, potentially sowing confusion and manipulate voters' perception about a candidate. It also does not adequately address other forms of misinformation and disinformation campaigns that may include false narratives about candidates, thereby failing to tackle deliberate attempts to undermine the democratic process. Additionally, the policy could further strengthen its stance on false claims related to election integrity, voter fraud, and the legitimacy of election results, which have been critical issues in recent elections.

- Regarding the digitally altered media, both the misinformation policy and manipulated media policy acknowledges that content is removed because "it can go viral quickly and experts advise that false beliefs regarding manipulated media often cannot be corrected through further discourse." However, it requires the media to be words-based video content crafted with advanced AI/ML tools, effectively excluding content manipulated using less sophisticated tools or conduct-based content.

- The impact of manipulated content on the political landscape is strikingly evident in recent instances. A case in point is a suspected deepfake video, reportedly created by an opposition party leader, depicting the Malaysian economic affairs minister (and potential successor to PM Mahathir Mohamad) in a sexual tryst with a party staffer. A mere suspicion of sodomy resulted in several arrests and nearly thwarted the succession plan. Another example involves the arrest of Jakarta's then-governor, an ethnic-Chinese Christian, after a video of campaign event where he said voters should not be swayed by those "using the Koran as a political tool" was edited to omit the word "use", which left plenty of room for ambiguity. During the 2019 Indonesian election, an online video depicted the seizure of millions of pre-marked ballot papers sent from China. Despite being debunked, the video had already been featured in approximately 17,000 tweets, creating doubts in the minds of voters. Current policies, particularly those addressing misinformation and manipulated media, are ill-equipped to effectively combat the rising tide of false and manipulated content during election campaigns. While the limitations of existing AI/ML tools may at present allow for the identification of manipulated content through discrepancies in, for instance, facial expressions and eye movements, this status quo is shifting. With technology evolving rapidly, the creation of hyper-realistic deepfakes, virtually indistinguishable from genuine content, is on the horizon. This poses an imminent threat to the integrity of democratic institutions and processes, demanding a more inclusive and comprehensive policy framework.

Meta's human rights responsibilities when it comes to video content that has been altered to create a misleading impression of a public figure, and how they should be understood with developments in generative artificial intelligence in mind.

As a preliminary matter, content moderation by social media intermediaries in general, and Meta in particular, relies on constitution-esque content policies, exogenous human rights instruments, and independent commitments as normative benchmarks.

- Although Meta does not have the obligations of governments under the International Covenant on Civil and Political Rights (ICCPR), their wide-ranging social and political impact necessitates them to assess the same kind of questions about protecting their users' right to freedom of expression. Previously, intermediaries moderated almost entirely without reference to the human rights implications. Now, Meta's Corporate Human Rights Policy, which serves as the foundation of the company's human rights commitments, reaffirms the company's commitments to the United Nations Guiding Principles on Business and Human Rights (UNGPs) and its interpretive guide as well as the ICCPR. These frameworks provide a principled and pragmatic model that is well-suited to the fast-paced, uncertain and complex landscape of the twenty-first century technological advancements. Commitments are implemented applying human rights policies and maintaining oversight, governance and accountability, prioritizing the most salient human rights issues in each context based on severity (scope, scale, remediability) and likelihood.

  Central to Meta's human rights responsibilities, therefore, is the application of Article 19 of the ICCPR, which recognises the right of every individual,without discrimination, to freedom of expression, which includes the "freedom to seek, receive and impart information and ideas *of all kinds*, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice." It includes political dissent, discourses, and commentary on public affairs, as well as offensive expressions and fake news. However, this right is not absolute and may be subject to certain restrictions imposed by law, provided it is necessary and proportionate with respect to the rights *or* reputations of others. Significantly, the chilling effect that the restrictive measures may have on expression and free flow of information necessitates its protection to be the rule, and the interference, properly justified, to remain an exception. Furthermore, Article 5(1) of the ICCPR restrains interpretation that allows actions that could destroy or excessively limit the recognized rights and freedoms. It is in these contexts that our assessment will elaborate Meta's human rights responsibilities.

  - With respect to content of political discourse, the value placed upon uninhibited expression is particularly high, and indeed expression considered insulting and offensive to a public figure (including heads of state and government) is insufficient to legally justify restriction. General comment No. 34 (on the right to freedoms of opinion and expression under Article 19 of the ICCPR) and General comment No. 25 (on participation in public affairs and the right to vote under Article 25 of the ICCPR) notes that the "free communication of information and ideas about public and political issues between citizens, candidates and elected representatives is essential," and must therefore be fully protected. Similarly, the Joint Declaration on Media Freedom and Democracy states that large online platforms should privilege "public interest content on their services in order to facilitate democratic discourse." A functioning democracy and freedom of expression are mutually reinforcing and complementary. However, restrictions may be imposed where the expression is not a legitimate criticism or political opposition, and instead immoderately attacks the reputation of the individual.

Here, the digitally altered content showing, on a loop, Biden placing a sticker on or around his granddaughter's chest with a suggestive song (containing the lyrics "girls rub on your titties"), accompanied by a caption calling him "a sick pedophile" for touching her breast, does not appear to be a genuine political critique or expression of dissent. It transgresses beyond the permissible limits of allowable insults that can reasonably be directed at a political leader, and instead creates misleading impressions about his character, family values, personal morals and intentions. It is a well-settled position of the Human Rights Committee that false portrayal of individuals, or deliberately spreading false rumors about them to generate public aversion, damages honour and reputation. It is also contrary to the recognitions in the preamble of the ICCPR that "inherent dignity and of the equal and inalienable rights of all members of the human family is the foundation of freedom, justice and peace in the world" and that "these rights derive from the inherent dignity of the human person." Given the visceral immediacy and virality of a short-form video and the ease of its dissemination using social media and encrypted messaging services to a potentially unlimited and illimitable audience, reinforced by the long-accepted truism that seeing is believing, it is likely to have greater impact on his reputation than, for instance, a still image. With more than 3.5 billion daily active users, Meta has reached a "scale of connectedness [that] is unprecedented in human history," and this only increases the severity and likelihood of the impact. Ahead of the election, this and other videos will resurface, and social media users will either amplify or attenuate their spread. As a result, in our view, the altered content is likely to constitute an impermissible personal attack on Biden, who, notwithstanding his standing as a public and political figure, is entitled to protection of his honour, reputation and dignity under the ICCPR.

○ Secondly, restrictions are allowed to safeguard rights of an individual, which includes human rights and more generally in international human rights law. In this context, Article 1 of the ICCPR confers the right of self-determination to every individual, so that they are freely able to determine their political status. This right has variously been described as the "right to authentic self-government, that is, the right of a people *really and freely* to choose its own political regime." Additionally, in relation to Article 25 of the ICCPR, General comment No. 25 asserts the importance of enabling individuals to *freely and independently* support or oppose government, and to vote, without undue influence, coercion, inducement, or *manipulative interference* of any kind.

Characterizing voters "mentally unwell" in this context seemingly draws a correlation between the conduct and his presidency, which could compromise his overall integrity, trustworthiness, respectability and credibility as a candidate for the 2024 presidential election, and therefore the ability of the voters to choose their next president freely and independently, without manipulative interference. Plausibly, this could be a part of broader misinformation campaign against Biden: several faux clips emerged showing him singing the opening lyrics of Baby Shark after announcing that he will sing the national anthem, publicly admitting he is old and may have dementia, using profanities and acknowledging that he knows he "was not [the voters'] first choice in 2020", acting disoriented and asking his wife whether he took his medicine, recommending troop deployment due to Russo-Ukraine war and the impending Chinese blockade of Taiwan, and admitting to getting his salary and pension paid in "in MILFs, orgies and top-tier f**king ice-cream flavors that will make your tiny little maggot lizard brain melt faster than the polar ice caps." Considering videos often have a much more immediate and powerful effect than traditional media, and individuals tend to accept video content at face value as evidence of truth, there is an increased susceptibility for the content to change perceptions of the voters and the outcome of the upcoming elections.

In this context, it is worth noting that it is well-documented that social media can influence the outcome of elections and events surrounding it. At a colloquium organized by UNESCO and the Global Network Initiative in 2018, the use of social media and technologies to spread misinformation, disinformation and hate speech during elections was recognized. For instance, a video was shown by President Recep Tayyip Erdoğan during a political rally of his main challenger, Kemal Kılıçdaroğlu, receiving an endorsement from a designated terrorist organization. In Slovakia, inauthentic content against the leader of the Progressive party Michal Šimečka, discussing vote-rigging, was released during a 48-hour moratorium ahead of the polls opening. Another video shows an advertisement by an opposition party in which Šimečka's voice has been used to say that he "used to believe in 70 genders and pregnant men." Ultimately, he lost the election. Malicious content that overwhelmed the 2020 US presidential race and seeded distrust about the legitimacy of Biden's victory culminated in the storming of the Capitol Building on 6 January by the supporters of then-President Donald Trump who believed his lies that the election was stolen from him. A survey by Brookings found that 57% of those surveyed have seen misinformation during the 2018 US elections and 19% believe it has influenced their vote. Hany Farid observed that altered media is resulting in stolen elections, which has "real-world consequences for individuals, for societies and for democracies." Hence, there is a strong argument to be made that the removal of altered media that creates a significantly misleading impression of Biden in particular, and public and political figures in general, is warranted.

- Aligned with Meta's voluntary human rights commitments, any restriction imposed must adhere to the three-part requirements of legality, legitimate aim, and necessity (and proportionality).

  - With respect to the principle of legality, Meta's existing policies on manipulated media, misinformation, adult nudity and sexual activity and coordinating harm and promoting crime appear insufficient and imprecise to effectively address the significant impacts on rights and reputation. For instance, the manipulated media policy creates an unnecessary dichotomy, by conditioning takedown for words-based video content crafted with advanced AI/ML tools that may mislead an ordinary person, and thereby excluding content manipulated using less sophisticated raster or vector graphics editors, or conduct-based inauthentic media. Similarly, the sexual activity policy is too constrictive, requiring a content to be "advertisements and recognised fictional images or with indicators of fiction [that shows] [s]queezing female breasts, defined as a grabbing motion with curved fingers that shows both marks and clear shape change of the breasts." Moreover, the misinformation policy fails to address manipulated media's impact on election integrity and outcomes, despite its stated objective to promote elections and commitment to "remove misinformation that is likely to directly contribute to a risk of interference with people's ability to participate in those processes." It is a fundamental requirement that the rules must "provide sufficient guidance to those charged with their execution … to enable them to ascertain what sorts of expression[s] are properly restricted and what sorts are not," so that these rules do not confer unfettered and arbitrary discretion to Meta and provide users with adequate guidance to enable them to regulate their conducts accordingly. It is also worth noting that under Article 2(2) of the ICCPR, Meta should take necessary steps to adopt clear and specific measures to address these policy gaps in alignment with the ICCPR.

    However, the bullying and harassment policy sets a more suitable tone by disallowing content with severe sexualized commentary and derogatory attacks. Distinction is drawn between public figures and private individuals, with content related to public figures only removed where the attacks are severe, taking into account relevant context and intent. Commitments to international human rights standards require contextual assessment of the relevant historical, political, linguistic and social nuances, as well as the context within which it was made, as content is not language- or context-agnostic. Thus, identification of actual and potential human rights impacts should start at a granular level, undertaking multi-faceted analyses of the specific user (or user category), as well as the geographic region and contexts in which use may lead to adverse impacts. Thus, the higher threshold for public figures notwithstanding, the incestuous connotation and unsubstantiated allegations of pedophilia against Biden, coupled with a sexually explicit song, is sufficiently severe to warrant removal of the at-issue content, especially considering Meta's intolerance towards such behavior and its stated position to "strive to create a more inclusive and equitable online environment for all users through our Community Standards and Community Guidelines, which prohibit hate speech, bullying, and harassment."

  - Any restriction on expression should pursue one of the legitimate aims listed in the ICCPR, including the rights or reputation of others. For reasons mentioned above, this requirement appears to be satisfied in this context.

  - Finally, the principle of necessity and proportionality provides that any restrictions on expression "must be appropriate to achieve their protective function; they must be the least intrusive instrument amongst those which might achieve their protective function; [and] they must be proportionate to the interest to be protected." Any restriction should substantiate in specific and individualized fashion the precise nature of the harm, and should ideally be limited to that specific content and not on the operation of entire sites and systems. Here, the significance of the internet and altered media in the context of the principle of necessity and proportionality is worth highlighting.

On a balance, we consider the removal of the video to serve the protective function of safeguarding honour, dignity and reputation of Biden (and his granddaughter) as well as the integrity of the elections, and a content-specific restriction to be the least intrusive and most proportionate response to the situation.