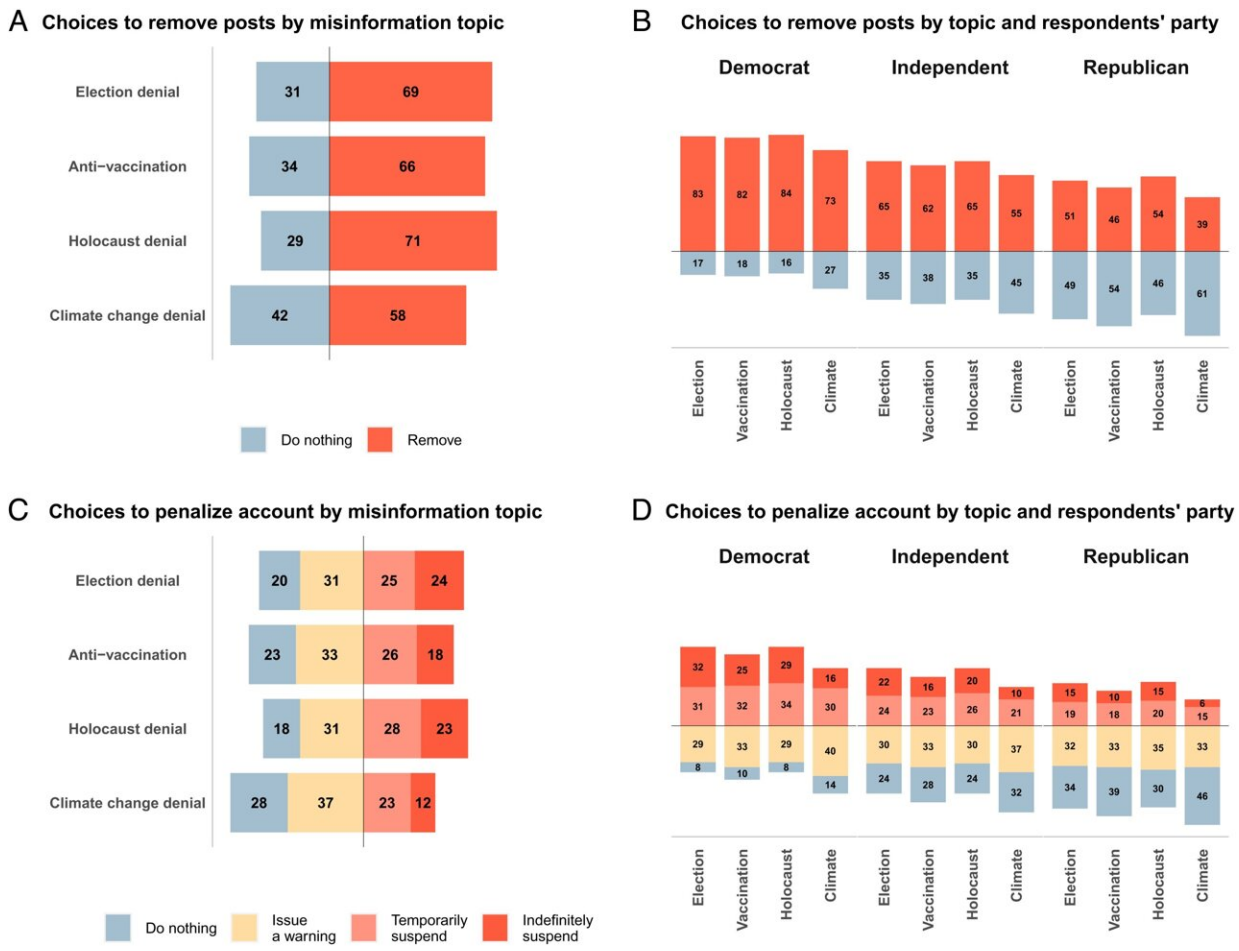
Free speech vs. harmful misinformation: How people resolve dilemmas in online content moderation

February 8 2023



Proportion of choices to remove posts and to suspend accounts. All numeric values represent percentages. (A) Choices to remove posts or do nothing by misinformation topic (all cases). (B) Choices to remove posts or do nothing, by topic and respondents' party affiliation. (C) Choices to penalize account by

misinformation topic (all cases). (D) Choices to penalize account by topic and respondents' party affiliation. N=40,845 cases evaluated in total. (Cases evaluated by Democrats, including Democrat-leaning, n=19,338; by independents n=8,229; by Republicans, including Republican-leaning, n= 13,278). Credit: *Proceedings of the National Academy of Sciences* (2023). DOI: 10.1073/pnas.2210666120

Online content moderation is a moral minefield, especially when freedom of expression clashes with preventing harm caused by misinformation. A study by a team of researchers from the Max Planck Institute for Human Development, University of Exeter, Vrije Universiteit Amsterdam, and University of Bristol examined how the public would deal with such moral dilemmas.

They found that the majority of respondents would take action to control the spread of misinformation, in particular if it was harmful and shared repeatedly. The results of the study can be used to inform consistent and transparent rules for content moderation that the general public accepts as legitimate.

The issue of content moderation on social media platforms came into sharp focus in 2021, when major platforms such as Facebook and Twitter suspended the accounts of then U.S. President Donald Trump. Debates continued as platforms confronted dangerous misinformation about the COVID-19 and the vaccines, and after Elon Musk singlehandedly overturned Twitter's COVID-19 misinformation policy and reinstated previously suspended accounts.

"So far, social media platforms have been the ones making key decisions on moderating misinformation, which effectively puts them in the position of arbiters of free speech. Moreover, discussions about online

content moderation often run hot, but are largely uninformed by empirical evidence," says lead author of the study Anastasia Kozyreva, Research Scientist at the Max Planck Institute for Human Development.

"To deal adequately with conflicts between free speech and harmful misinformation, we need to know how people handle various forms of moral dilemmas when making decisions about content moderation," adds Ralph Hertwig, Director at the Center for Adaptive Rationality of the Max Planck Institute for Human Development.

In the conjoint survey experiment, more than 2,500 U.S. respondents indicated whether they would remove social media posts spreading misinformation about democratic elections, vaccinations, the Holocaust, and climate change. They were also asked whether they would take punitive action against the accounts by issuing a warning or a temporary or indefinite suspension. Respondents were shown information about hypothetical accounts, including political leaning and number of followers, as well as the accounts' posts and the consequences of the misinformation they contained.

The majority of respondents chose to take some action to prevent the spread of harmful misinformation. On average, 66% of respondents said they would delete the offending posts, and 78% would take some action against the account (of which 33% opted to "issue a warning" and 45% chose to indefinitely or temporarily suspend accounts spreading misinformation). Not all misinformation was penalized equally: Climate change denial was acted on the least (58%), whereas Holocaust denial (71%) and election denial (69%) were acted on most often, closely followed by anti-vaccination content (66%).

"Our results show that so-called free-speech absolutists such as Elon Musk are out of touch with public opinion. People by and large recognize that there should be limits to free speech, namely, when it can

cause harm, and that content removal or even de-platforming can be appropriate in extreme circumstances, such as Holocaust denial," says co-author Stephan Lewandowsky, chair in cognitive psychology at the University of Bristol.

The study also sheds light on the factors that affect people's decisions regarding content moderation online. The topic, the severity of the consequences of the misinformation, and whether it was a repeat offense had the strongest impact on decisions to remove posts and suspend accounts. Characteristics of the account itself—the person behind the account, their partisanship, and number of followers—had little to no effect on respondents' decisions.

Respondents were not more inclined to remove posts from an account with an opposing political stance, nor were they more likely to suspend accounts that did not match their political preferences. However, Republicans and Democrats tended to take different approaches to resolving the dilemma between protecting free speech and removing potentially harmful misinformation. Democrats preferred to prevent dangerous misinformation across all four scenarios, whereas Republicans preferred to protect free speech, imposing fewer restrictions.

"We hope our research can inform the design of transparent rules for content moderation of harmful misinformation. People's preferences are not the only benchmark for making important trade-offs on content moderation, but ignoring the fact that there is support for taking action against misinformation and the accounts that publish it risks undermining the public's trust in content moderation policies and regulations," says co-author Professor Jason Reifler from the University of Exeter.

"Effective and meaningful platform regulation requires not only clear and transparent rules for content moderation, but general acceptance of

the rules as legitimate constraints on the fundamental right to free expression. This important research goes a long way to informing policy makers about what is and, more importantly, what is not acceptable user-generated content," adds co-author Professor Mark Leiser from the Vrije Universiteit Amsterdam.

More information: Anastasia Kozyreva et al, Resolving content moderation dilemmas between free speech and harmful misinformation, *Proceedings of the National Academy of Sciences* (2023). DOI: 10.1073/pnas.2210666120

Provided by Max Planck Society