ONLINE APPENDICES FOR:

Catalyst of hate? Ethnic insulting on YouTube in the aftermath of terror attacks in France, Germany, and the United Kingdom 2014-17

Forthcoming at *Journal of Ethnic and Migration Studies*

Christian S. Czymara (Goethe University Frankfurt)
Stephan Dochow-Sondershaus (Freie Universität Berlin)
Lucas G. Drouhot (Utrecht University, The Netherlands)
Müge Simsek (University of Amsterdam, The Netherlands)
Christoph Spörlein (Heinrich-Heine-Universität Düsseldorf)

Appendix A – Additional figures

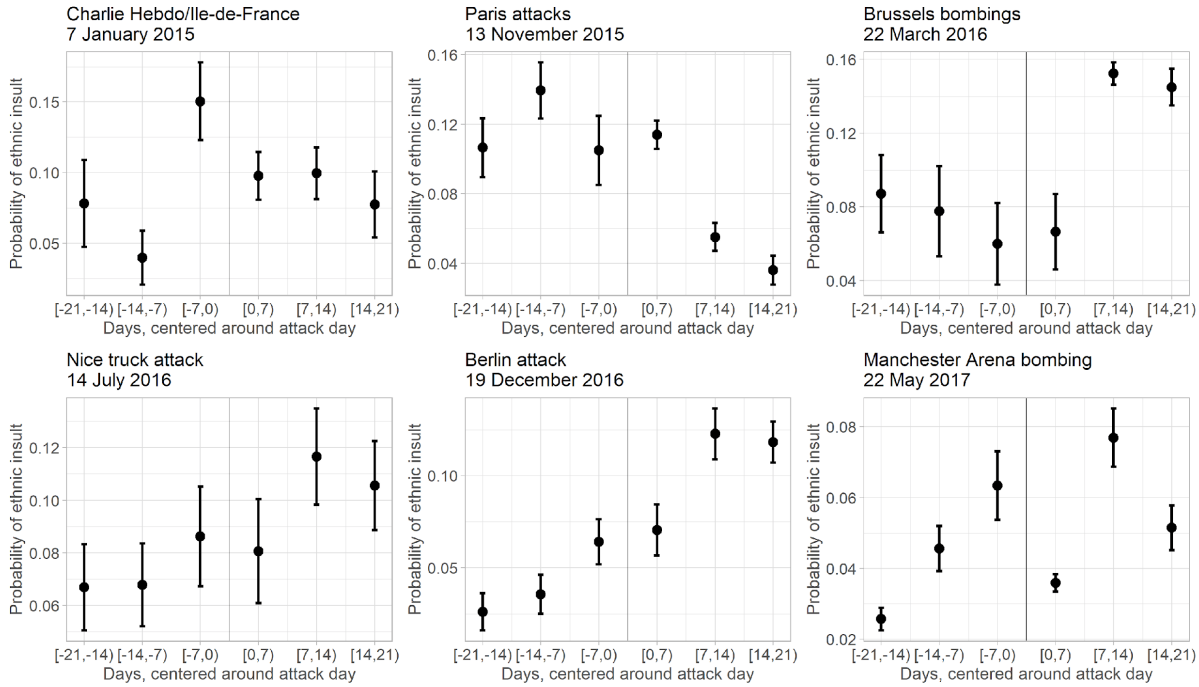

Figure A1. Linear probability models predicting the probability of online comments to contain ethnic insults in the three weeks before and after attacks (Charlie Hebdo/Ile-de-France, N_{obs} = 4055; Paris attacks, N_{obs} = 14832; Brussels bombings, N_{obs} = 20644; Nice truck attack, N_{obs} = 5878; Berlin attack, N_{obs} = 10389; Manchester Arena bombing, N_{obs} = 46598).

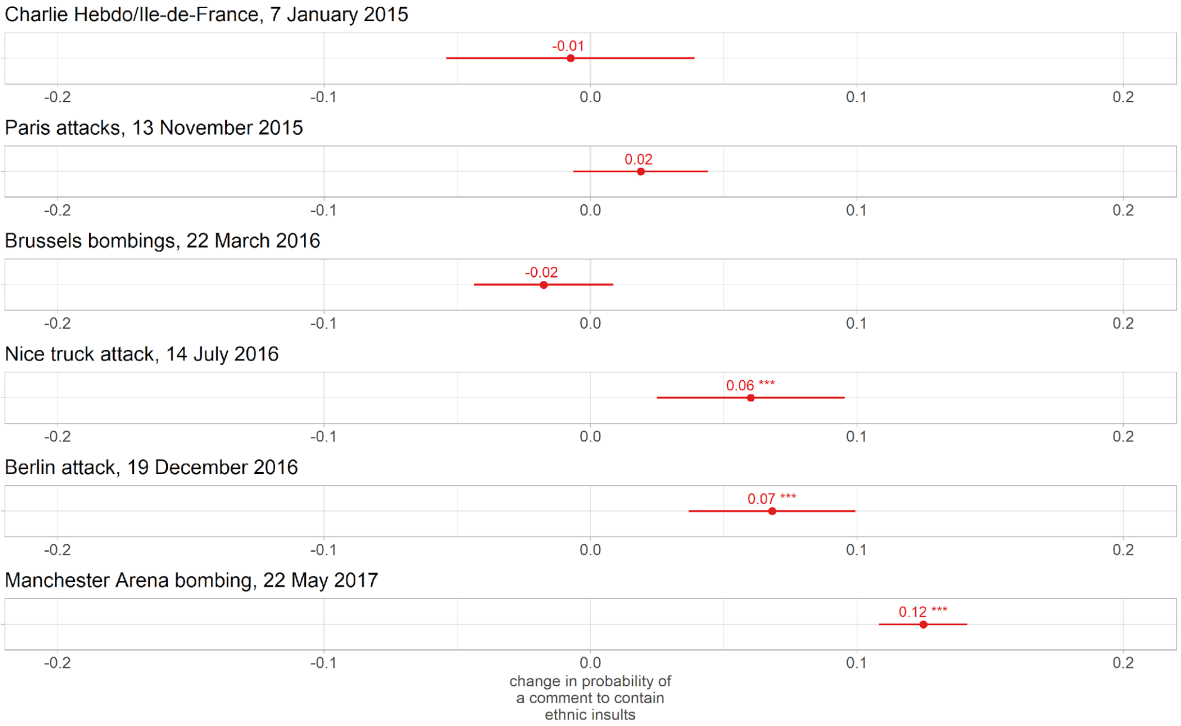

Figure A2 Linear probability models predicting the change in the probability of online comments to contain ethnic insults before and after attacks in Germany (Charlie Hebdo/Ile-de-France, $N_{obs} = 1310$; Paris attacks, $N_{obs} = 4425$; Brussels bombings, $N_{obs} = 16380$; Nice truck attack, $N_{obs} = 1874$; Berlin attack, $N_{obs} = 2669$; Manchester Arena bombing, $N_{obs} = 7094$).

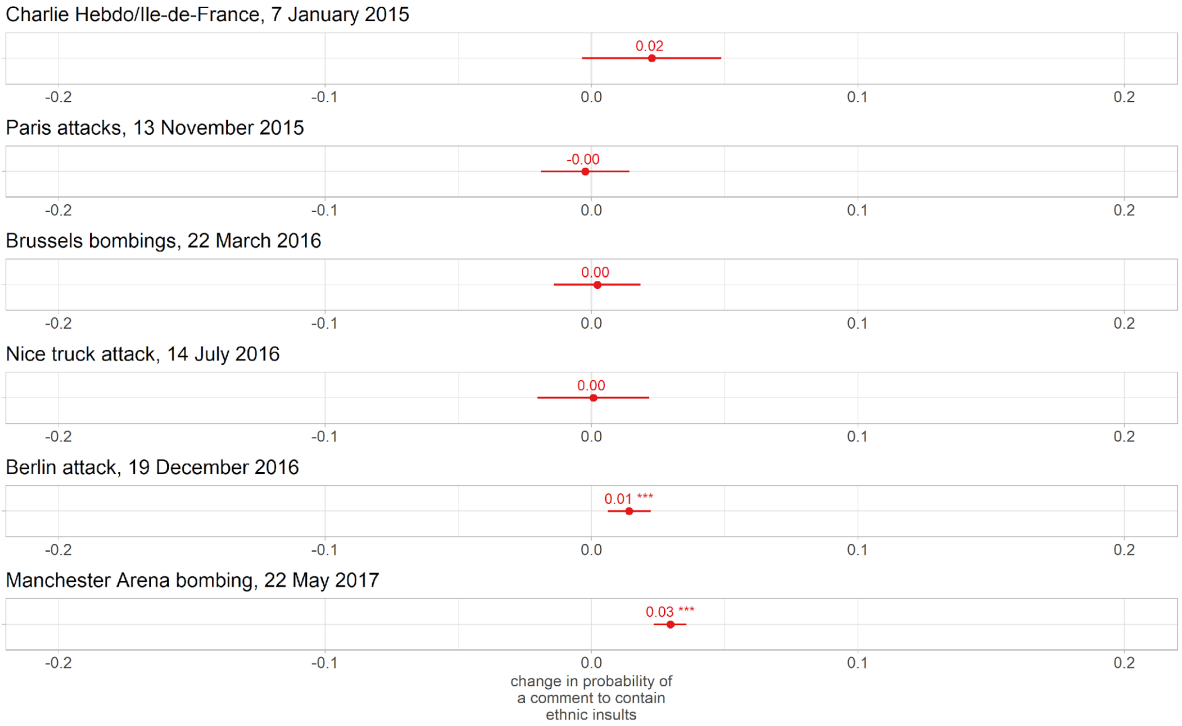

Figure A3 Linear probability models predicting the change in the probability of online comments to contain ethnic insults before and after attacks in France (Charlie Hebdo/Ile-de-France, N_{obs} = 1190; Paris attacks, N_{obs} = 2257; Brussels bombings, N_{obs} = 1352; Nice truck attack, N_{obs} = 944; Berlin attack, N_{obs} = 3617; Manchester Arena bombing, N_{obs} = 13365).

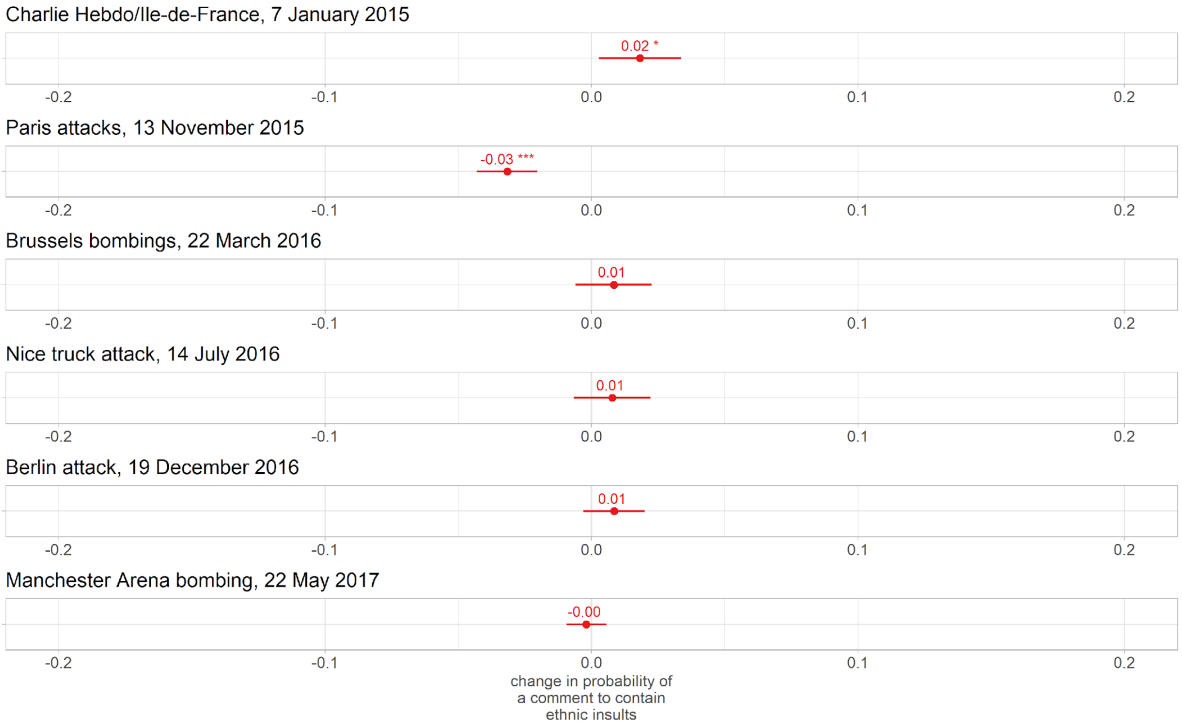

Figure A4 Linear probability models predicting the change in the probability of online comments to contain ethnic insults before and after attacks in the UK (Charlie Hebdo/Ile-de-France, N_{obs} = 1555; Paris attacks, N_{obs} = 8150; Brussels bombings, N_{obs} = 2912; Nice truck attack, N_{obs} = 3060; Berlin attack, N_{obs} = 4103; Manchester Arena bombing, N_{obs} = 26139).

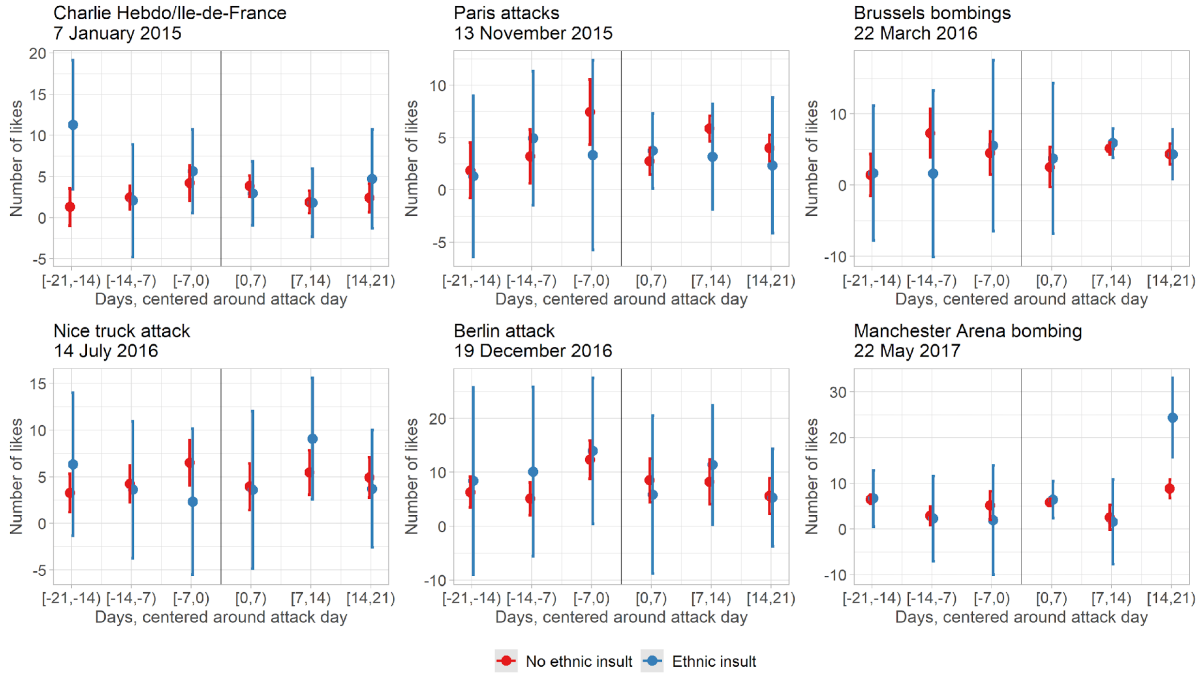

Figure A5. Predicted number of likes given to comments in the three weeks before and after attacks (Charlie Hebdo/Ile-de-France, N_{obs} = 4055; Paris attacks, N_{obs} = 14832; Brussels bombings, N_{obs} = 20644; Nice truck attack, N_{obs} = 5878; Berlin attack, N_{obs} = 10389; Manchester Arena bombing, N_{obs} = 46598).

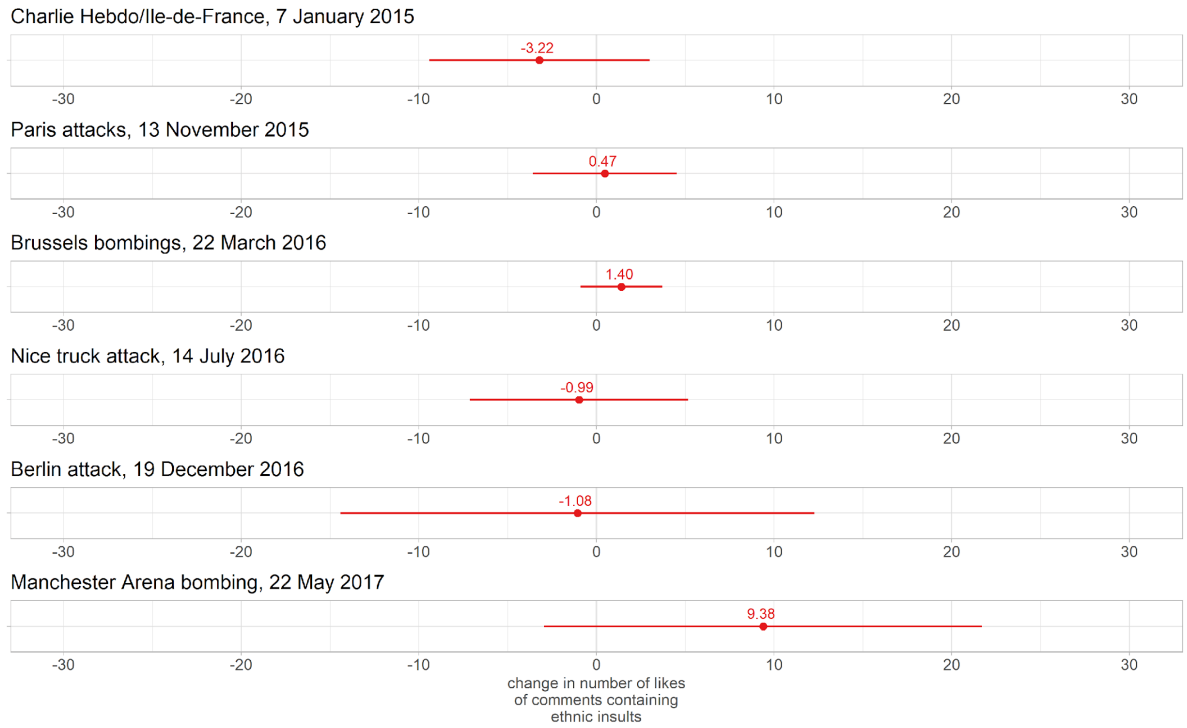

Figure A6. OLS models predicting the change in the number of likes of comments before and after attacks in Germany (Charlie Hebdo/Ile-de-France, $N_{obs} = 1310$; Paris attacks, $N_{obs} = 4425$; Brussels bombings, $N_{obs} = 16380$; Nice truck attack, $N_{obs} = 1874$; Berlin attack, $N_{obs} = 2669$; Manchester Arena bombing, $N_{obs} = 7094$).

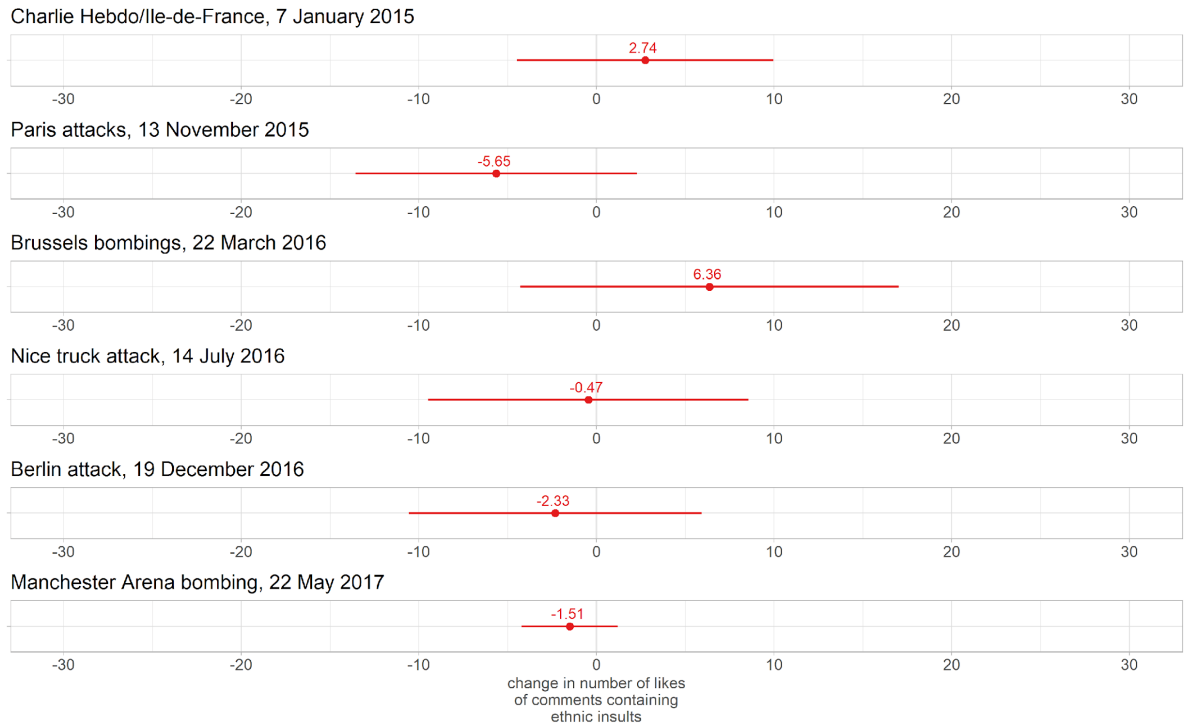

Figure A7. OLS models predicting the change in the number of likes of comments before and after attacks in France (Charlie Hebdo/Ile-de-France, N_{obs} = 1190; Paris attacks, N_{obs} = 2257; Brussels bombings, N_{obs} = 1352; Nice truck attack, N_{obs} = 944; Berlin attack, N_{obs} = 3617; Manchester Arena bombing, N_{obs} = 13365).

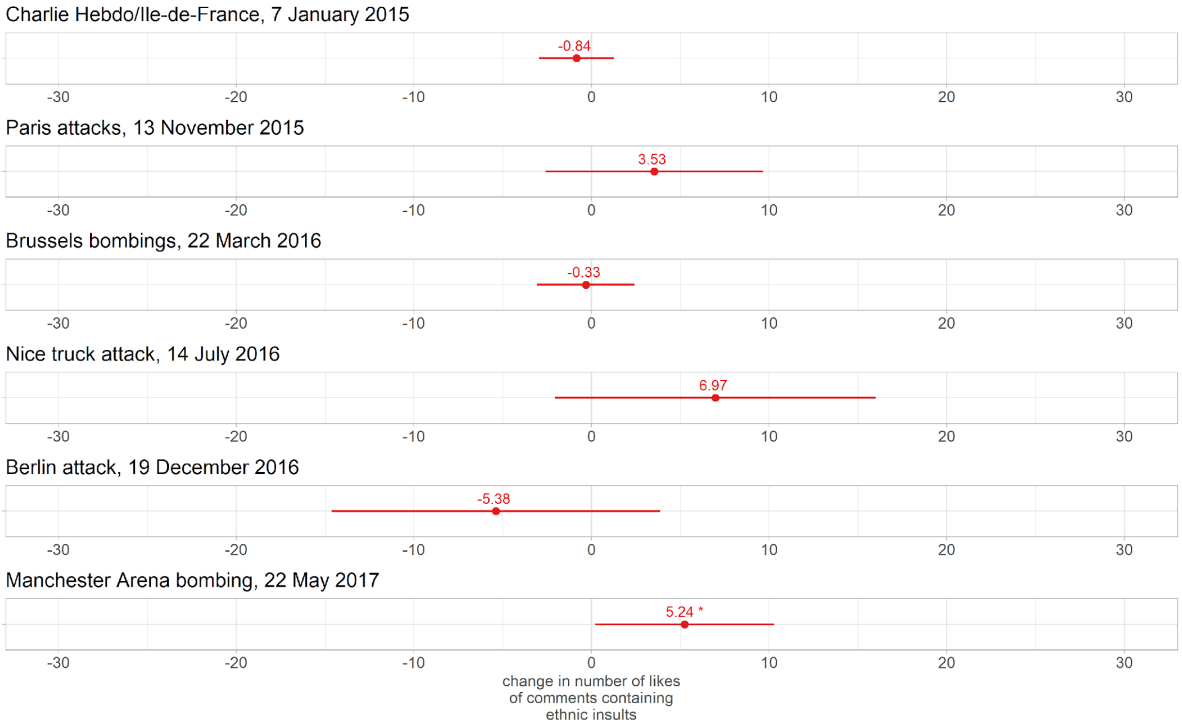

Figure A8. OLS models predicting the change in the number of likes of comments before and after attacks in the UK (Charlie Hebdo/Ile-de-France, N_{obs} = 1555; Paris attacks, N_{obs} = 8150; Brussels bombings, N_{obs} = 2912; Nice truck attack, N_{obs} = 3060; Berlin attack, N_{obs} = 4103; Manchester Arena bombing, N_{obs} = 26139).

Appendix B – Measuring classifier performance

Table B1. Classifier performance before and after Berlin Christmas market attack

	DE	UK	FR
True negative (before)	43	41	49
False negative (before)	0	7	0
False positive (before)	5	0	1
True positive (before)	1	2	0
True negative (after)	39	41	45
False negative (after)	2	5	1
False positive (after)	1	3	2
True positive (after)	8	1	2

The numbers are based on three hand-coded random samples of 50 comments before the attack and 50 comments after. By and large, the error rate of the classifier does not seem to differ before and after the attack. For example, there are 15 ethnic insults in the hand-coded data for the UK. Of these 15 insults, nine took place before the attack and six after the attack. Of the nine before the attack, the algorithm identified two correctly, missed seven but also did not falsely classify a non-insult as an insult. Of the six insults after the attack, the algorithm found one and classified three comments as insults that were not coded as such by the human coders. One of the three comments the algorithm wrongly classified as an ethnic insult reads "you fucking racist son of a bitch". This clearly is an insult, but not a racist one and rather one *against* a racist. However, these differentiations are hard to understand for a machine. Similarly, an example of an ethnic insult the algorithm missed is "brussels is importing low intelligence inbred people to take the place of europeans. wake up". This is easily detected as an ethnic insult for a human, but the seemingly neutral words lead the machine to classify the comment not clearly enough as an ethnic insult. Thus, while the algorithm is not perfect in finding the hand-coded ethnic insults, this does not seem to depend on the timing relative to the attack. This implies that the foundation for an unbiased estimation of the attack effect, the assumption of temporal exogeneity, does not seem to be violated.

Appendix C – Identification of insults using a dictionary approach

As an alternative to Perspective API (machine-based classifier approach), we created custom language-specific dictionaries to identify insults targeted at minorities. Specifically, we used a dictionary for generally insulting words (e.g., 'fuck'), one for ethnic identifiers (e.g., 'Muslim') and one for ethnophaulisms (e.g., 'towelheads'). According to this approach, comments were labelled as ethnically insulting if they contained a) both common insults and ethnic identifiers or b) ethnophaulisms. We reran our main analyses using the insult classifier variable obtained by this dictionary-based approach. Results from the alternative analyses (Figures C1-C3) are mostly in agreement with the results presented in the main text (Figure 2, Figure 4, and Figure 5). The main differences are found with respect to the effects of attacks on the probability of a comment to contain ethnic insults. In particular, we find conflicting effects for the Brussel bombings (0.07 in the machine-based approach vs -0.08 in the dictionary-based approach), for the Nice truck attacks (0.03 in the machine-based approach vs null effect in the dictionary-based approach) and for the Charlie Hebdo shootings (null effect in the machine-based approach vs 0.05 in the dictionary-based approach). There is also one minor difference between the two approaches regarding the like counts of insulting comments: the null coefficient for the Manchester Arena bombing in the machine-based approach becomes significant and positive in the dictionary-based approach.

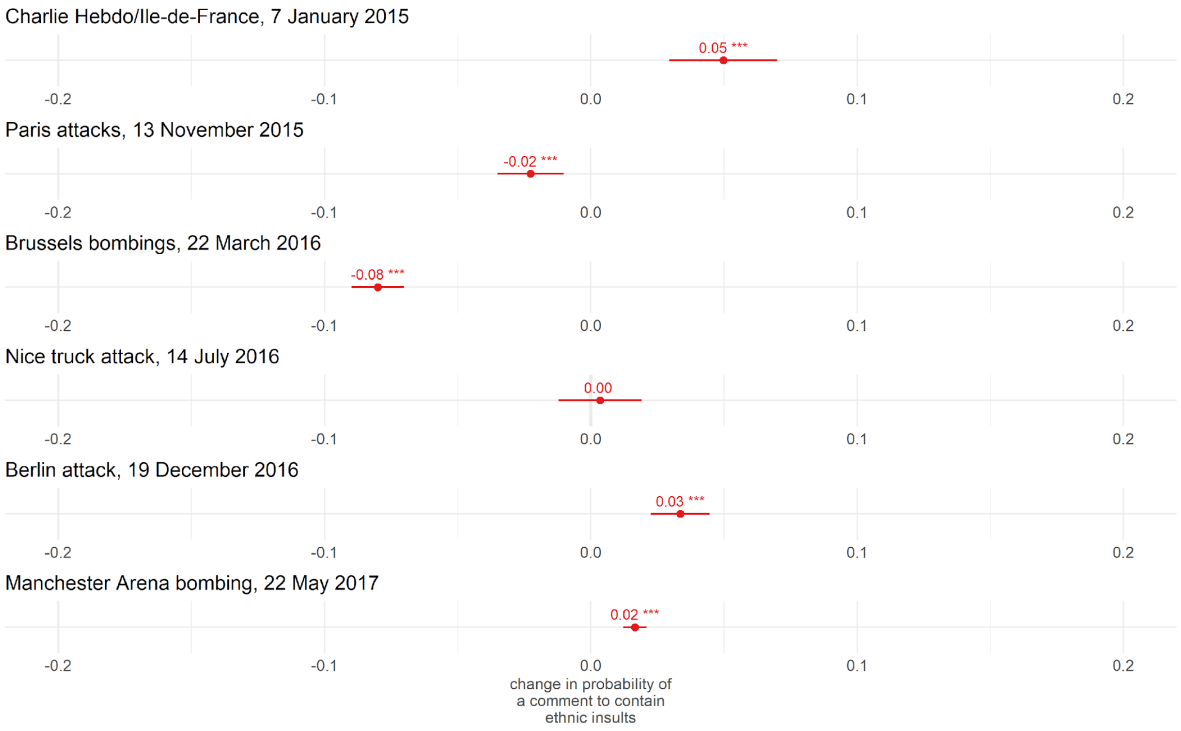

Figure C1. Linear probability models predicting the change in the probability of online comments to contain ethnic insults before and after attacks (Charlie Hebdo/Ile-de-France, N_{obs} = 4055; Paris attacks, N_{obs} = 14832; Brussels bombings, N_{obs} = 20644; Nice truck attack, N_{obs} = 5878; Berlin attack, N_{obs} = 10389; Manchester Arena bombing, N_{obs} = 46598).

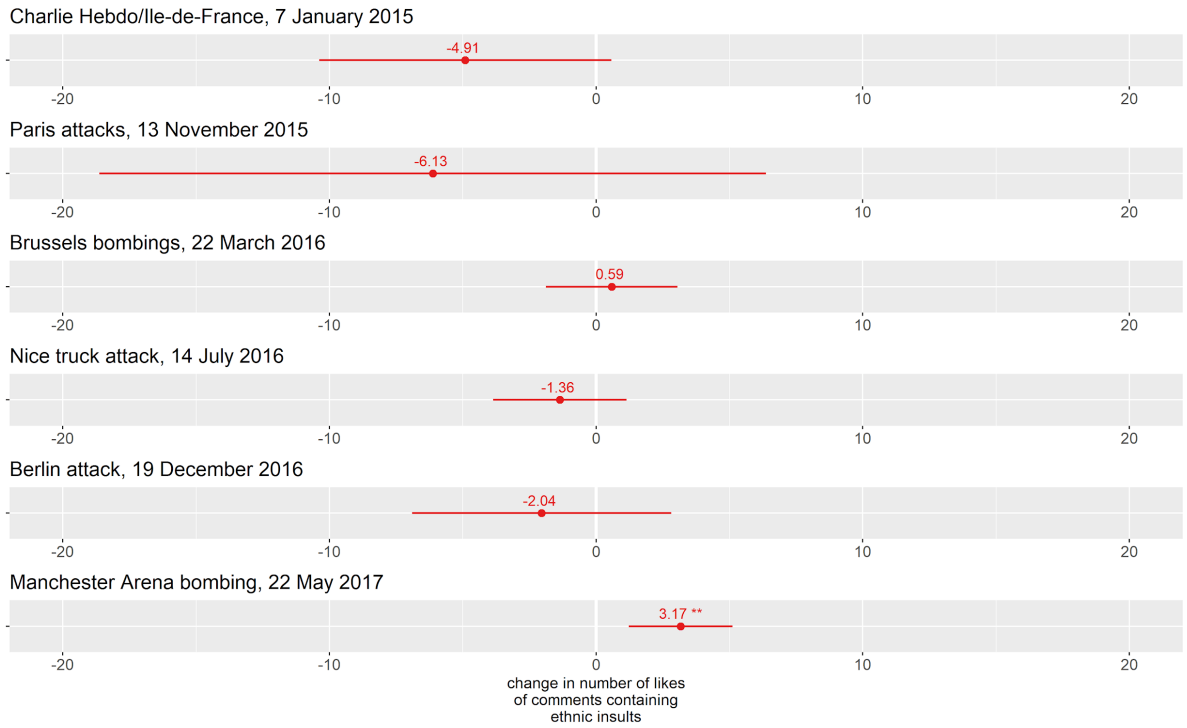

Figure C2. OLS models predicting the change in the number of likes of comments before and after attacks (Charlie Hebdo/Ile-de-France, N_{obs} = 4055; Paris attacks, N_{obs} = 14832; Brussels bombings, N_{obs} = 20644; Nice truck attack, N_{obs} = 5878; Berlin attack, N_{obs} = 10389; Manchester Arena bombing, N_{obs} = 46598).

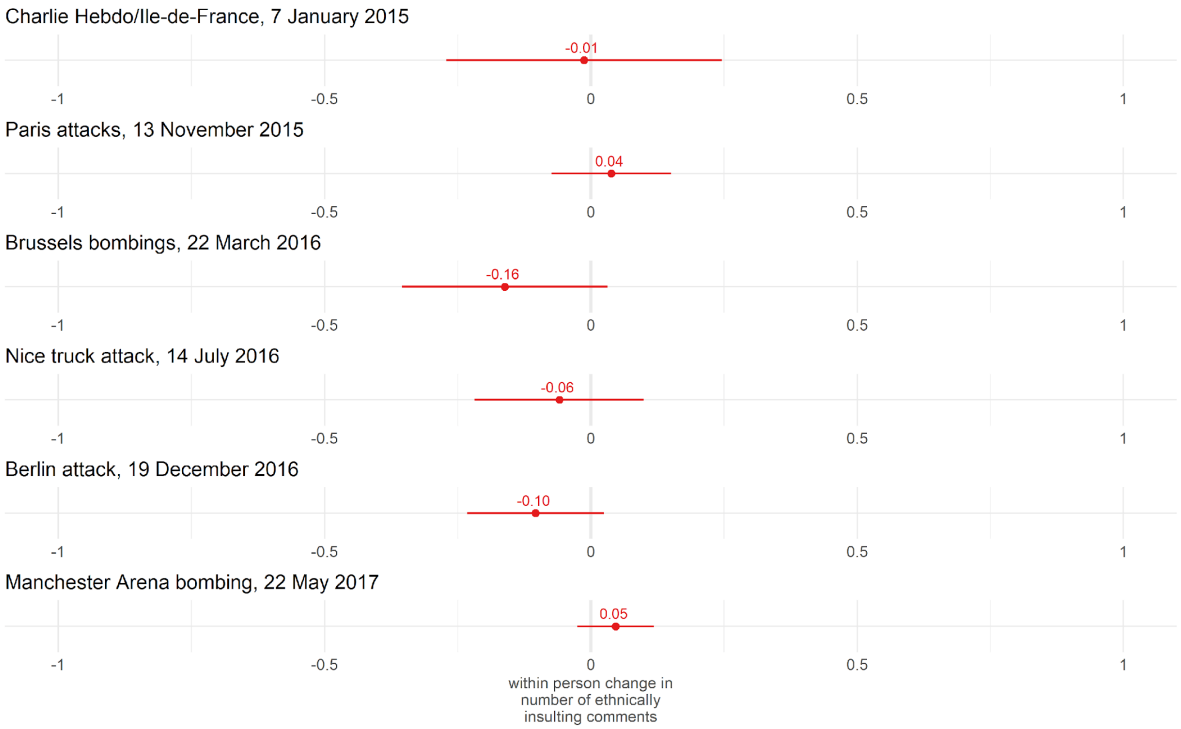

Figure C3. Linear models with user fixed-effects predicting the within-person change in the number of ethnic insults before and after each attack (Charlie Hebdo/Ile-de-France, $N_{obs} = 156$; Paris attacks, $N_{obs} = 472$; Brussels bombings, $N_{obs} = 334$; Nice truck attack, $N_{obs} = 236$; Berlin attack, $N_{obs} = 500$; Manchester Arena bombing, $N_{obs} = 992$).

Appendix D – Additional video statistics

Manually examining titles and content of videos that received 100 or more comments leads us to the conclusion that our videos are generally very mixed and there are no outright racist or insulting videos per se within this subsample (but there are critical videos about Muslims or immigration). In particular, the videos that received more than 100 comments are a mix of:

- Broad audience channels (of famous YouTubers, including political commentators)
- Videos that were originally broadcasted by public news media outlets (often reposted by users of different political views) consisting mostly of the fragments of news, political talks and debates, and media interviews.
- Snippets of everyday video footage, for example from demonstrations, etc.

Overall, however, most videos received only a small number of comments (only one comment within the chosen time window is very common), whereas some viral videos received a large share of comments. This is reflected in a skewed distribution of the number of comments per video with a median of about 5 and a mean of about 55 comments per video. Furthermore, Figure D1 shows that in the majority of time-windows for each country, there are more ethnically insulting comments after the attack in videos that were posted after the attack. However, these videos make up only a small proportion of our overall video space from which we have data on the comments. This is shown in Figure D2, which suggests that the overwhelming majority of videos from which we extracted comment data were already present before the event window. Taking these statistics together suggests that single famous YouTube videos act as rallying points that attract many commentators, including hateful ones, after a terror attack.

Figure D1. Number of ethnic insults *after the attack* per video by the date of the first comment that a video received.

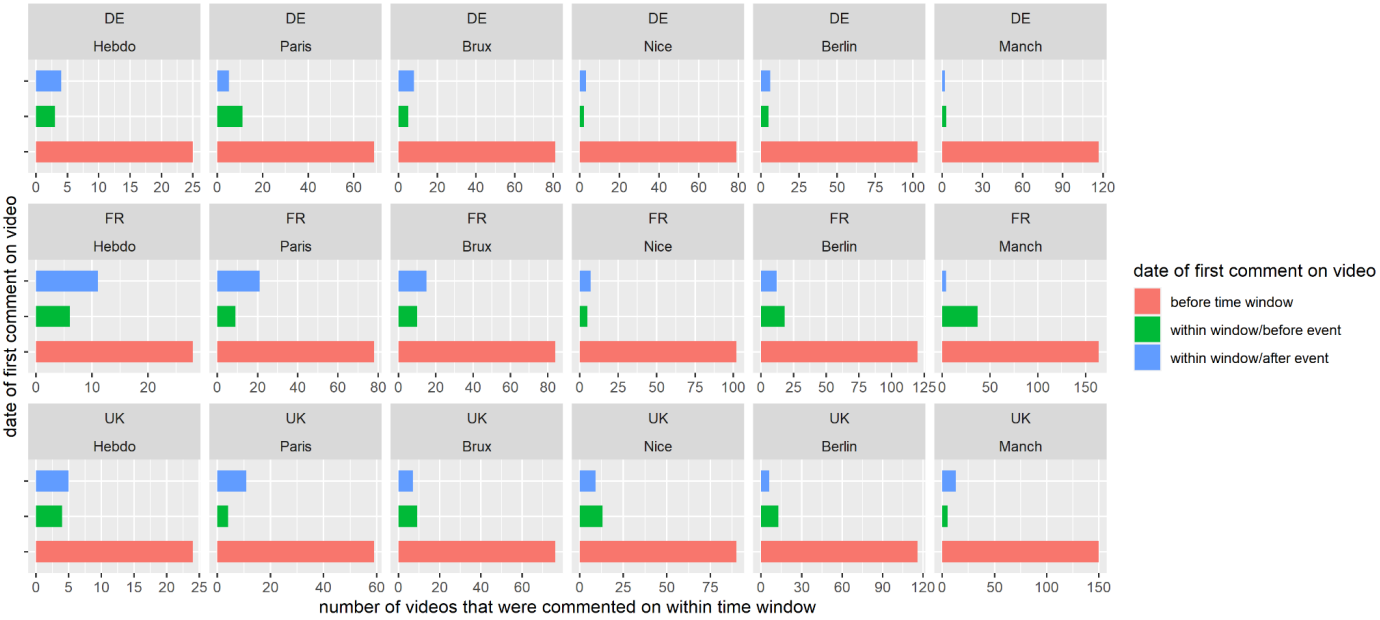

Figure D2. Number of videos by the date of the first comment that a video received.

There are single videos that were posted after an attack and attracted a lot of comments. Some of these popular videos were related to the attack, for example, Rayk Anders, a political commentator posted an influential German video about the Manchester attack after the event. Other videos were not thematically connected to terror attacks, such as Jan Böhmermann's video "BE DEUTSCH", which is a satire on national identity and was posted after the Brussels attack.