Journal of Affective Disorders 294 (2021) 769–775

Contents lists available at ScienceDirect

Journal of Affective Disorders

journal homepage: www.elsevier.com/locate/jad

Research paper

What's in a trauma? Using machine learning to unpack what makes an event traumatic

Payton J. Jones *

Department of Psychology, Harvard University, 33 Kirkland St. #1240, Cambridge, MA 02138, United States

ARTICLE INFO

Keywords:
Trauma
Judgment
PTSD

ABSTRACT

What differentiates a trauma from an event that is merely upsetting? Wildly different definitions of trauma have been used in both formal (psychiatric) and informal (cultural, colloquial) settings. Yet there is a dearth of empirical work examining the features of events that individuals use to define an event as a 'trauma.' First, a group of qualitative coders classified features (e.g., actual physical injury, loss of possessions) of 600 event descriptions (e.g., "was verbally harassed by a boss," "watched a video of an adult being shot and killed"). Next, across two studies, machine learning was used to predict whether individuals rated event descriptions as 'trauma' or 'traumatic' in over 100,000 judgment tasks. In Study 1, examining continuous ratings from 'not at all traumatic' to 'extremely traumatic,' a cross-validated LASSO regression with polynomial features provided the best out-of-sample predictions ($r^2 = 0.76$), outperforming ridge regression, support vector regression, and linear regression. In Study 2, using binary judgments, a random forest model accurately predicted out-of-sample individual responses ($AUC = 0.96$), outperforming a neural network and an AdaBoost ensemble classifier. The most important event features across the two studies were actual death, threat of death, and the presence of a human perpetrator. The most important human features in predicting judgments were political orientation and gender.

1. Introduction

What makes people call certain events 'trauma' or 'traumatic?' This study found that judgments about whether an event is a 'trauma' can be predicted with about 90% accuracy. Events were most consistently seen as traumatic when they contained strong signals of current or future threats to survival ('threat of death,' 'actual death') and when they were rated by females and those of more liberal political orientation.

What differentiates traumatic events from events that are merely upsetting or negative? This question has haunted psychiatrists attempting to formulate diagnostic criteria for posttraumatic stress disorder (PTSD). The original definition of a traumatic event in the DSM-III and DSM-III-R denoted an event "outside the range of normal human experience" that would "be markedly distressing to almost anyone" (APA, 1987, p. 250; see also APA, 1980, pp. 236-238). DSM-IV markedly expanded the definition of trauma (notably, by including learning about events rather than directly witnessing or experiencing them; APA, 1994). While there were several solid empirical reasons for expansions (see McNally, 2015), the expanding definition of trauma caused alarm amongst some researchers. These researchers expressed concern that expanding the breadth of 'trauma' would undermine the integrity of the psychobiological concept of PTSD (Bracha & Hayashi, 2008; Elhai, Kashdan, & Frueh, 2005; McNally, 2003; McNally, 2009). The definition of trauma was somewhat scaled back in DSM-5 (Pai, Suris, & North, 2017).

Outside of formal psychiatric epidemiology, the expansion of 'trauma' extends far beyond the DSM definition. For example, psychologist Monnica Williams (2015) points to the "traumatizing role" of microaggressions and asserts that learning secondhand about negative events happening to one's racial group can result in "vicarious traumatization." Others have identified news stories as a source of trauma (Lees, 2018); in the wake of the Kavanaugh allegations, the New York Times lead with the headline *When the News Itself is a Form of Trauma* (Jacobs, 2018). Others have increasingly emphasized the role of "intergenerational trauma" and "historical trauma," which refer to traumas that are not directly experienced (or sometimes never even known to the victim) but nevertheless cause emotional wounds centuries later (Coyle, 2014).

Haslam (2016) convincingly argues that a whole cluster of harm-related concepts (e.g., bullying, abuse) have similarly expanded.

* Corresponding author.
E-mail address: paytonjjones@gmail.com.

https://doi.org/10.1016/j.jad.2021.07.066
Received 28 April 2021; Received in revised form 13 July 2021; Accepted 15 July 2021
Available online 27 July 2021

He notes that such expansions can be either 'horizontal' (when the concept grows to encompass qualitatively new phenomena) or 'vertical' (when the severity threshold is lowered). The horizontal and vertical expansions of trauma are not necessarily a bad thing. The term trauma derives from the Greek for 'wound' and was used to refer to physical injuries long before the emergence of the PTSD concept (this original definition persists in phrases such as 'traumatic brain injury;' Haslam, 2016). The horizontal emergence of trauma as a psychological concept has surely enriched the science and treatment of psychopathology. Vertically expanding concepts of harm may reflect greater sensitivity, reducing uncertainty about the unacceptability of certain actions and empowering victims to take action (Cikara, 2016).

Regardless of where (or if) one draws a bright line to delineate traumas from non-traumas, it is useful to understand how laypersons make decisions about which events are traumatic. The purpose of this study is to understand the extent to which the objective features of events (e.g., whether the event involved physical pain, whether the event involved a human perpetrator) can be used to predict whether individuals will call an event 'trauma' or 'traumatic.' A panel of individuals used qualitative coding to identify objective features of 600 descriptions of events (e.g., "was scolded by parents," "was socially excluded," "was raped by a close friend," "watched a TV show in which an adolescent died by suicide"). Other objective features were coded through rudimentary natural language processing (e.g., a *witness* category for each description containing the stems 'witness*' or 'watch*').

In Study 1, these features are used to predict the mean ratings of event descriptions on a Likert Scale ranging from 1 (Not at all traumatic) to 7 (Extremely traumatic). After dividing the data into a training and test set, a simple linear regression and cross-validated LASSO were fit to identify which features are most important in predicting mean ratings of events. A cross-validated ridge regression model and support vector regression were then fit to estimate the degree to which trauma ratings in the holdout data can be accurately predicted.

In Study 2, the event features were combined with participant information to predict more than 100,000 binary participant judgments across two experimental studies. In this case, the participants were asked to classify events as either 'trauma' or 'not trauma'. After dividing the data into a training and test set, three cross-validated models (neural network [multilayer perceptron], random forest, and AdaBoost ensemble learning) were used to predict individuals' classifications. The classification accuracy in the test set was then evaluated using receiver operator characteristic (ROC) analysis. The feature importance was then assessed using permutation importance, Gini importance, and visualized using partial dependence plots.

2. Study 1

2.1. Method

2.1.1. Qualitative coding of events

Six individuals qualitatively coded the 600 event descriptions in terms of 14 features. First, coders met and were trained on the codebook using an initial set of items that was pre-coded by the researcher (items 1-100; directed content analysis, Hsieh & Shannon, 2005). The researcher and coders discussed and reached consensus on any necessary amendments to the item codes, amendments to the codebook, or new domains needed to address the breadth of possible objective features. Items were discussed until saturation was reached (i.e., a consensus was reached that the codebook contained the breadth and detail needed to code all events unambiguously). Coders met twice more, jointly coding new sets of items as a training exercise (items 100-115; items 116-216), again discussing and amending the codebook until saturation was reached. Finally, all qualitative coders independently coded the remaining items using the finalized codebook as a guide (items 217-600). Final codes for the items were derived using majority rule. Rater consistency metrics were calculated exclusively from the items that were rated independently (items 217-600).

2.1.2. Procedure

Participants were recruited from Amazon Mechanical Turk (MTurk). Participants were eligible for the study if there were adult United States residents and had an MTurk approval rating of at least 95%. Participants were immediately excluded if they failed a CAPTCHA or a US residency screener (e.g., "What emergency number is most common in the United States?"). Three attention checks were interspersed throughout the task (e.g., "If you are reading this question, please select 1"). Participants were excluded if they incorrectly answered any attention checks. Non-demographic measures (i.e., the stimuli) were unique to this study, reducing concerns about non-naivete to measures common on crowdsourcing platforms (Chandler et al., 2015; Robinson et al., 2019). After applying exclusion criteria, 250 participants remained.

We randomly divided the event descriptions into six sets of 110 items each. Among these 110 items, 12 items were a constant subgroup that appeared in all six sets, whereas the other 98 were unique to their set. The descriptions in a given set were presented in random order to participants ($n_{total} = 250$, $n_{set} \approx 42$), who were asked to rate each event description on a 7-point Likert scale ranging from "Not at all traumatic" to "Extremely traumatic." Participants then reported their demographic information.

2.1.3. Analysis

To verify the consistency of raters in different sets, we calculated interrater reliability on the consistent set of 12 items given to all participants. We then calculated the mean rating for each event description by taking a simple average across all participants who rated the event ($n_{set} \approx 42$). This mean rating was the dependent variable for our primary analysis in Study 1 ($n = 600$). The 14 event features from the coders were used as predictors in addition to the 3 features drawn from text analysis. The data were divided into training (80%) and test data (20%). The analysis was conducted in Python, and models were fit using the scikit-learn library (Pedregosa et al., 2011). Full analysis code is available in the supplemental materials (https://osf.io/n3p6g/).

Four separate models were used: simple linear regression, L1 penalized LASSO regression, ridge regression, and support vector regression. Each model serves a distinct purpose related to our research aims. The simple linear regression is useful because the parameters are easily interpretable compared to the other models. For instance, a parameter value of 1 for the *threat of death* feature would mean that, holding other features constant, the 1-7 Likert ratings were 1 point higher for events that were coded as containing a threat of death. The linear regression also serves as a point of comparison for the more complex models.

What about interactions between features? Including all pairwise interactions leads to a large number of parameters, and in the simple linear regression this might lead to overfitting or overinterpretation. The LASSO model is useful because it provides a penalty that helps shrink small parameters to zero. If we hope to interpret interaction parameters, LASSO can help us avoid overfitting. One downside of LASSO is that eliminating features entirely (shrinking their parameter to zero) might be ideal for interpretation, but it might decrease the predictive performance. Ridge regression also involves adding a penalty to avoid overfitting, but parameters are typically retained rather than being driven to zero. In most cases one would therefore expect ridge regression to have superior predictive performance. Support vector regression is even more flexible in detecting nonlinear patterns and combinations of variables. These models were included to investigate the extent to which trauma ratings were predictable based on the event features

3. Results

3.1. Qualitative coding

Coder reliability and agreement was calculated based on the events that were coded independently (items 217-600). The full results are available in Table 1. Seven of the categories demonstrated excellent agreement (>90%), six demonstrated good agreement (>75%), and one category had poor agreement. In terms of Krippendorf's Alpha reliability, five of the categories fell in the excellent range (>0.8), four in the acceptable range (>0.66), and the remaining five categories in the poor range. Only one category had both poor agreement and reliability (threat to moral worldview). Because the qualitative codes were a means to an end (prediction of trauma ratings), rather than an end themselves (interpretation of individual item codes), all categories were retained in the primary analyses even when category reliability was suboptimal. One possible limitation is that poor reliability may cause certain categories to have less predictive power, and so their importance may be underestimated. To assess the scope of this limitation, the category reliability and agreement were tested for correlations with the feature importance of categories, and these correlations are reported in sections concerning feature importance.

3.2. Participants

Participants had a mean age of 36.1 (sd = 10.5). Participants identified as male (55%), female (43%) or other (2%). Participants identified as Caucasian/White (77%), Asian/Pacific Islander (6%), Black/African American (5%), Hispanic (5%), Native American/Alaska Native (1%), or multiracial (6%). 12% of the sample identified as Hispanic, 13% reported a past diagnosis of one or more mental disorders, and 31% reported experiencing a Criterion A trauma in their lifetime. Participants leaned slightly liberal in their political orientation ($mean$ = 3.4 on a scale ranging from 1 = *Very conservative* to 5 = *Very liberal*). A majority of participants (55%) were nonreligious.

Table 1
Qualitative Coding: Reliability and Agreement by Category, KA = Krippendorf's alpha. Rather than being coded by human raters, Learned about, Witness, and Child were algorithmically coded categories based on whether word stems (e.g., learn*) appeared in the description.

Category	KA	Agreement	Example Item
Threat of physical injury	0.62	81.0%	had a tire explode while driving
Actual physical injury	0.83	92.7%	lost an eye in a kitchen accident
Threat of death	0.76	88.3%	was shot while working as a police officer
Actual death	0.92	94.8%	witnessed a friend being beaten to death
Sexual content	0.92	97.9%	was asked for sex by a boss
Physical pain	0.72	76.6%	lost a foot due to frostbite
Human perpetrator	0.80	80.5%	was tackled by a security officer
Close interpersonal perpetrator	0.82	94.3%	was slapped in the face by a mother
Close interpersonal victim	0.75	92.7%	learned about a family member being beheaded
Loss of possessions	0.84	98.4%	lost all life savings due to fraud
Threat to moral identity	0.59	92.2%	accidentally injured a child while drunk
Threat to social status or other identity	0.65	87.2%	lost significant language function after a stroke
Threat to moral worldview	0.43	65.9%	witnessed a child being shot and killed
Threat to trust in interpersonal relationships	0.56	87.8%	was molested by a father
Learned about (learn*, read*)	NA	NA	learned about the murder of a family member
Witness (witness*, watch*)	NA	NA	witnessed a son being forcibly arrested
Child (child*)	NA	NA	killed a child pedestrian while driving

3.3. Interrater reliability

In terms of rating the items on the scale from 'not at all traumatic' to 'extremely traumatic,' interrater reliability across each of the 250 participants as separate judges was good (ICC_1 = 0.70; Shrout & Fleiss, 1979). When considering the mean rating for each of the 12 constant items across each set, interrater reliability was excellent (i.e., six judges; ICC_1 = 0.99). That is, each group of raters was highly consistent in how they rated the 12 constant items.

3.4. Prediction

In the training data, the cross-validated performance of the LASSO regression, ridge regression, and support vector regression outperformed the linear regression (r^2 = 0.80, 0.80, 0.78, 0.72). The same general pattern was replicated in out-of-sample prediction of the holdout test data (r^2 = 0.76, 0.75, 0.75, 0.70). This suggests that interactions between the features meaningfully aided prediction. Figure 1 shows the coefficient values for each of the 17 primary features given to the models (i.e., interactions excluded). A supplemental table shows the coefficient values of interaction terms that were nonzero in the LASSO model (https://osf.io/n3p6g/). Some of the notable interactions with positive coefficients included *close interpersonal victim + learned about, threat to social status or other identity + learned about*, and *threat to moral worldview + child*. To check for problems with rater reliability, coefficient values were tested for correlation with category reliability and agreement. The correlations were nonsignificant in all cases, though consistently positive in direction (r = 0.16-0.40, p = 0.16-0.69).

3.5. Study 1 discussion

In Study 1, the objective features of event descriptions were used to predict individual's ratings of those descriptions on a Likert scale ranging from 'not at all traumatic' to 'extremely traumatic.' A cross-validated LASSO model provided the best out-of-sample performance, followed closely by a ridge regression and support vector regression, all of which outperformed a simple linear model. Model performance was strong overall, explaining 70-76% of the variance in ratings in the test set.

Some of the most consistently important features were *threat of death, actual death, threat to social status or other identity*, or involved a *close interpersonal victim*. Interestingly, these results seem consistent with an evolutionary perspective, as these categories seem to represent clear threats to reproductive fitness (in contrast with some surprisingly less predictive items, including *physical pain* and *threat to moral worldview*). An alternative explanation is that some of the most predictive items are relatively unambiguous markers of a major event (there is little ambiguous about a death), whereas the less predictive categories were more ambiguous. For instance, many descriptions involved relatively minor physical pain (i.e., "twisted an ankle"), and *threat to moral worldview* was one of the least reliable categories among the qualitative coders.

4. Study 2

4.1. Method

4.1.1. Context

Study 2 utilized data that were originally collected in two experiments conducted by Jones et al., 2021. The procedures and materials of those experiments are documented in detail elsewhere (https://osf.

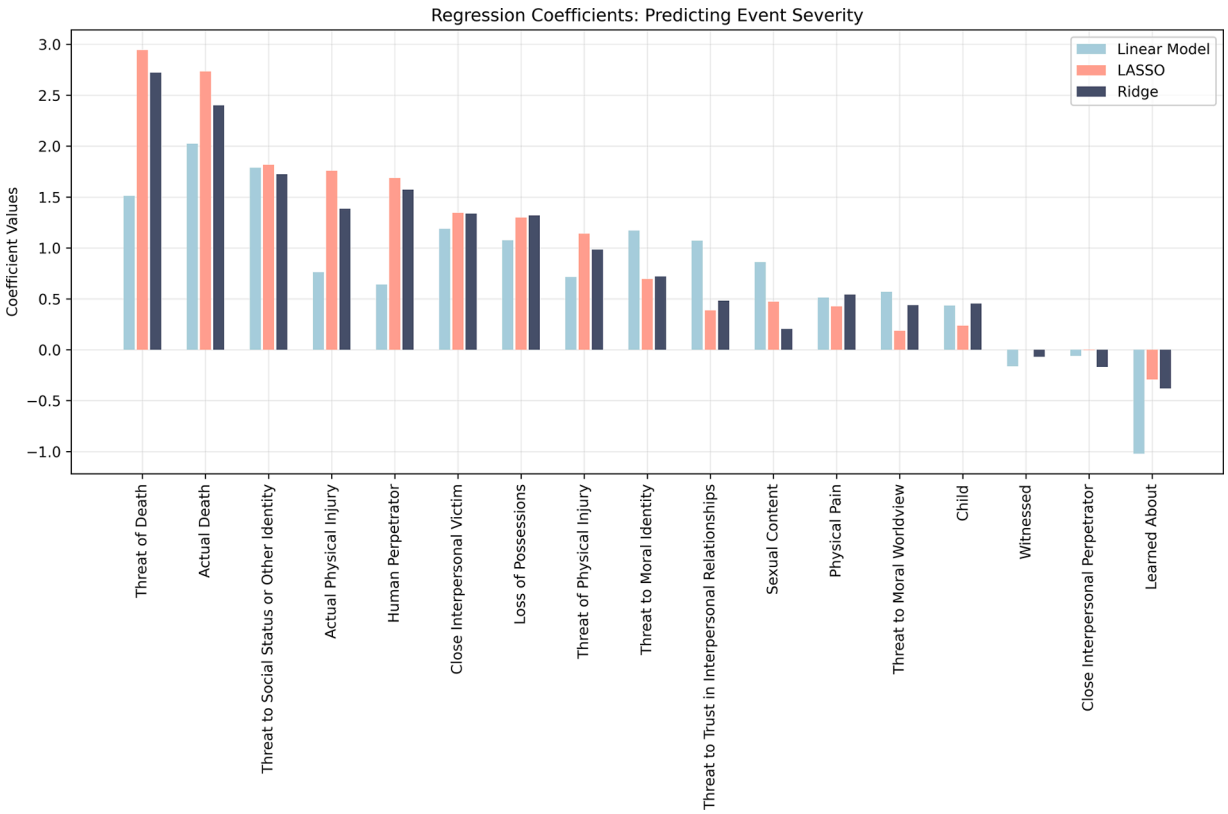

Fig. 1. This figure shows coefficient values for each event feature in the linear model, LASSO, and ridge regression. Threat of death and actual death had consistently high coefficient values across models.

io/3e2us/). The event descriptions seen by participants were a subset of the event descriptions in Study 1. Event descriptions were intentionally selected to have acceptably low standard deviations based on the Study 1 data ($sd < 1.6$) to avoid choosing items with overly ambiguous interpretations.

4.1.2. Procedure

Participants were recruited from MTurk, and were screened with the same inclusion and exclusion criteria as in Study 1. Participants from Study 1 were excluded from participating in Study 2. After full exclusion criteria (i.e., attention checks) were applied to those initially qualifying for the survey, a total of 543 participants remained across both experiments ($n = 276, 267$).

Participants were sequentially shown descriptions of events ($p = 300, 90$) and were asked to rate them in a binary manner as either 'trauma' or 'not trauma.' As these data were collected as part of a separate experiment, participants were randomized into two experimental groups that influenced which events they were shown and in what order. The experimental conditions are not relevant to the present study except that the assigned condition is included as a control variable in all analyses.

4.2. Analysis

In Study 1, the goal was to predict the mean rating of an event description given objective features of that item. In other words, the sample was the total set of event descriptions ($n = 600$). In Study 2, the goal is to predict the exact response ('trauma,' 'not trauma') of a given participant seeing a given item. That is, the sample is each classification made by a participant ($n = 300*276 + 90*267 = 106{,}830$).

This setup allows for additional inputs into our model. For instance, it is now possible to include features of the individual participants as predictors. Gender, race, ethnicity, religiosity, political orientation, and age were included as predictors. Categorical variables were dummy coded. The trial number in which they saw the description (i.e., the order) was also included.

Three separate models were fit to the data: a neural network (multilayer perceptron), random forest, and AdaBoost ensemble learning model. The intention in fitting various distinct models was to identify a model that maximally predicts (out-of-sample) participant responses. All models were initially fit with cross-validation in a training set to select optimal tuning parameters. The best model was then selected via the area under the curve (AUC) in a receiver-operator characteristic (ROC) analysis fit on a holdout test set.

After selecting the best model, feature importances were calculated using a permutation feature importance approach. The permutation importance is calculated by selecting one feature at a time, randomly shuffling the data in that feature, and examining the degree to which the model prediction degrades as a result. This analysis was performed in the holdout test set. Patrial dependence plots were also generated and are provided in the supplemental materials (https://osf.io/n3p6g/).

5. Results

5.1. Participants

Participants had a mean age of 37.0 ($SD = 11.1$). Participants identified as male (57%), or female (43%). Participants identified as Caucasian/White (80%), Black/African American (7%), Asian/Pacific Islander (5%), Hispanic (3%), or multiracial (6%). 6% of the sample identified as Hispanic, 16% reported a past diagnosis of one or more mental disorders, and 30% reported experiencing a Criterion A trauma in their lifetime. Participants leaned slightly liberal in their political orientation (*mean* = 2.5 on a scale ranging from 1 = *Very liberal* to 5 = *Very conservative*). A majority of participants (56%) were nonreligious.

5.2. Prediction

All models performed well in both the training and test data. In the training data, the random forest model appeared to perform best, followed by the multilayer perceptron (MLP) and AdaBoost models (*AUC* = 0.99, 0.96, 0.95; *prediction accuracy* = 0.96, 0.90, 0.88). This same pattern was replicated in the test data with the random forest model narrowly outperforming the multilayer perceptron and AdaBoost model (*AUC* = 0.96, 0.96, 0.95; *prediction accuracy* = 0.90, 0.90, 0.88). A figure displaying the ROC curve for the random forest model is available in the supplementary materials (https://osf.io/n3p6g/).

Feature importances can also be generated. Beyond comparing to Study 1, additional features (of the participants) can be examined. Figure 2 shows the permutation importances for each of the features in the model. Because the random forest model was chosen as the best fitting model, impurity-based feature importances (Gini importances) were available. These are also presented in Figure 2 (normalized by their maximum value). As a note, the Gini importance can be misleading in some cases as it is biased towards categories with many options (as can be seen in its high values for trial number and age). As such, the permutation importance is preferred in this scenario. Correlations between feature importances and category reliability and agreement were inconsistent in direction and nonsignificant (r = -0.42-0.35, p = 0.14-0.59).

6. Discussion

All three models performed well in predicting participant judgments of 'trauma' or 'not trauma,' with a prediction accuracy around 90% in the holdout data. This suggests that we have captured much of the relevant feature space when it comes to determining which features of descriptions make the descriptions appear traumatic and which features of individuals predispose them to rating certain kinds of events as traumatic or not traumatic.

As in Study 1, actual death, threat of death, and physical injury emerged as important predictors. As mentioned earlier, this fits with an evolutionary lens; perhaps events that affect survival are paramount in human judgments of trauma and harm. One feature that emerged as especially important in Study 2 is whether the event involved a human perpetrator. This is consistent with a robust literature suggesting that interpersonal violence is more generative of PTSD compared to non-interpersonal events, such as accidents and natural disasters (Liu et al., 2017; Kessler et al., 2017; Shalev et al., 2019). Other features, such as threats to moral worldview and loss of possessions, may be comparatively less important. The importance of the experimental conditions (included as control variables) re-confirms the analysis provided by Jones et al., 2021: exposure to a higher range of event descriptions in an experiment did indeed influence the degree to which participants rated events as traumatic.

Interestingly, participant features were overall less important for predicting outcomes, suggesting consistency in ratings across different types of groups. For instance, race was an unimportant factor in predicting judgments. Of the participant features, the most impactful were political orientation (conservatives were less likely to rate events as 'trauma'), gender (males were less likely to rate events as 'trauma'), and age (younger people were less likely to rate events as 'trauma').

Previous studies have noted that women are more likely to develop PTSD following trauma (e.g., Tolin & Foa, 2006; Breslau & Anthony, 2007; Chung & Breslau, 2008). This is partly due to the disproportionate impact of rape and sexual assault, which cause PTSD at high rates and can sensitize individuals to future traumas (Breslau & Anthony, 2007), but even after controlling for these factors, women have higher conditional rates of PTSD across all trauma types (Tolin & Foa, 2006). In other words, women are both more likely to rate an event as 'trauma' and to suffer PTSD following that event. One possible explanation is that when deciding whether an event qualifies as a 'trauma,' individuals rely on

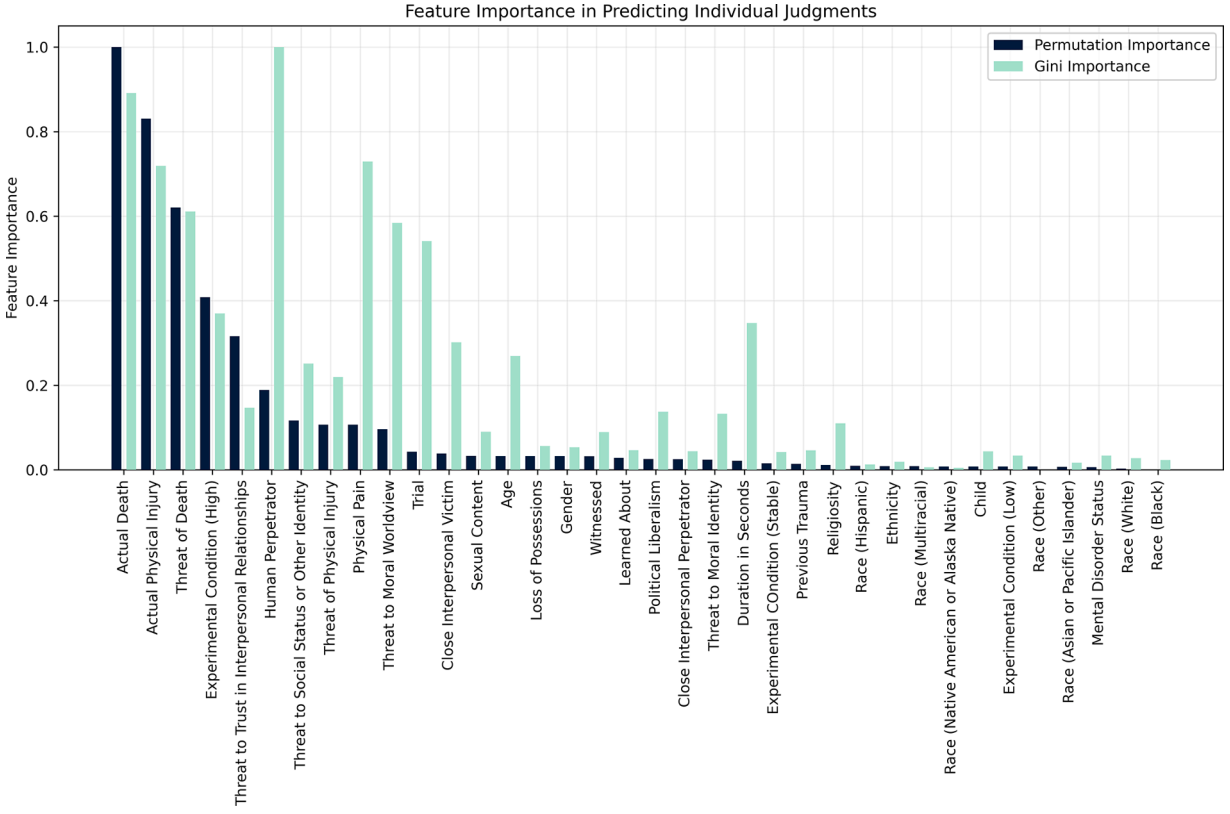

Fig. 2. This plot shows normalized permutation importance and Gini importance for each event feature and participant feature in the random forest model (the best performing model in the holdout data). Event features were overall more important than participant variables.

(accurate) perceptions of their own vulnerability to that event. It's also possible that the perception of whether an event is 'trauma' actively shapes trauma outcomes, or that another risk factor such as anxiety sensitivity influences both ratings of events and vulnerability to PTSD (Jones & McNally, 2021; Berntsen & Rubin, 2015).

The finding on age is interesting because it contradicts the notion that individuals in younger cohorts have more expanded views of trauma. Findings on age and broadened concepts of harm have indeed been mixed (McGrath, Randall-Dzerdz, Wheeler, Murphy, & Haslam, 2019; Jones & McNally, 2021). This suggests that the increasing breadth of trauma over time (i.e., Haslam, 2016; Vylomova, Murphy, & Haslam, 2019) may be due to a period effect rather than a cohort effect.

6.1. Limitations

There are several limitations to these studies. Our sample was restricted to adults using Amazon Mechanical Turk (MTurk). Although our results suggested that ratings were relatively consistent across demographic groups, recruitment from MTurk means that certain individuals (those without access to the internet, children) are not represented, limiting the generalizability of the studies. In addition, the stimuli themselves may not be representative of the total set of possible event descriptions. That is, there might be relevant features of events that are absent from the stimulus set or from the coding scheme. Further, these potential missing features could be importantly related to participant variables that were deemed unimportant in the current study.

Qualitative coding was used to capture supposedly objective features of events (i.e., *actual death, physical pain*). Although there was good agreement across raters for most categories, there were some categories (*threat to moral worldview*) where agreement was poor. Although reliability did not significantly correlate with feature importance, the direction of this correlation was typically positive had a medium effect size in the case of some models. It's also possible that human biases might have made coders more likely to select certain codes when the event appeared more subjectively 'traumatic' to them. One future direction might be training natural language processing models to classify the objective features to determine whether the same level of prediction can be achieved without using human coders. This study only tested a limited range of models and model parameters. Other models may provide superior out-of-sample prediction compared to the models presented in this paper.

This analysis also captures participants at a static point in time. Yet diachronic analyses suggest that the semantics of 'trauma' are rapidly changing over time (Haslam & McGrath, 2020; Vylomova et al., 2019). Thus, the present results may only speak to a very specific era of human history and in a specific population (US-resident adults).

6.2. Conclusion

Definitions of trauma have been a contentious point in psychiatric classification and in the broader sociopolitical landscape. Trauma and related concepts (bullying, abuse) have expanded over time, altering how we deal with certain types of negative events (Haslam, 2016; Cikara, 2016). This manuscript provides one of the first data-driven attempts to understand how individuals make decisions about whether specific events qualify as 'trauma.'

Across two studies, findings suggested that participants' judgments of potentially traumatic events are predictable based on objective features of the event. The most important features across studies were actual death, threat of death, and the presence of a human perpetrator. Other features, such as whether the event posed a potential threat to one's moral worldview, whether the event involved a child, or whether the event involved a close interpersonal figure as a perpetrator, mattered less. Of the available participant characteristics, political liberalism, female gender, and older age predicted higher likelihood of classifying an event as a trauma. Overall, features of the events mattered much more than the characteristics of the participants who were rating them, indicating high agreement on what makes an event a 'trauma.'

Contributors

P.J. designed the study, collected the data, directed qualitative coding, conducted the analyses, and wrote the manuscript. The author has approved the final version of the article.

Role of funding

This project was supported by the National Science Foundation (Grant Number DGE1745303). The funder had no role in study design, collection, analysis, or interpretation of the data. The funder had no role in the writing of the report or the decision to submit the article.

Declaration of Competing Interest

The author declares no conflict of interest.

Acknowledgments

Silvana Gomez, Melissa Guineau, Claire Hotchkin, Nicole Iannella, Noah Ramos, and Margot Steinberg were qualitative coders for the 600 event descriptions used in the study.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jad.2021.07.066.

References

Berntsen, D., Rubin, D.C., 2015. Pretraumatic stress reactions in soldiers deployed to Afghanistan. Clin. Psychol. Sci. 3, 663–674.
Bracha, H.S., Hayashi, K., 2008. Torture, culture, war zone exposure, and posttraumatic stress disorder criterion A's bracket creep. Arch. Gen. Psychiatry 65, 115–116.
Breslau, N., Anthony, J.C., 2007. Gender differences in the sensitivity to posttraumatic stress disorder: An epidemiological study of urban young adults. J. Abnorm. Psychol. 116 (3), 607.
Chandler, J., Paolacci, G., Peer, E., Mueller, P., Ratliff, K.A., 2015. Using nonnaive participants can reduce effect sizes. Psychol. Sci. 26 (7), 1131–1139.
Cikara, M., 2016. Concept expansion as a source of empowerment. Psychol. Inquiry 27, 29–33.
Chung, H., Breslau, N., 2008. The latent structure of post-traumatic stress disorder: tests of invariance by gender and trauma type. Psychol. Med. 38 (4), 563–573.
Coyle, S., 2014. Intergenerational trauma: Legacies of loss. Social Work Today 14 (3), 18.
Elhai, J.D., Kashdan, T.B., Frueh, B.C., 2005. What is a traumatic event? Br. J. Psychiatry 187, 189–190.
Haslam, N., 2016. Concept creep: Psychology's expanding concepts of harm and pathology. Psychol. Inquiry 27, 1–17.
Haslam, N., McGrath, M.J., 2020. The creeping concept of trauma. Soc. Res. Int. Quart. 87 (3), 509–531.
Hsieh, H.F., Shannon, S.E., 2005. Three approaches to qualitative content analysis. Qual. Health Res. 15 (9), 1277–1288.
Jacobs, J., 2018. When the news itself is a form of trauma. New York Times. Retrieved from. https://www.nytimes.com/2018/09/26/us/metoo-survivors-kavanaugh-cosby.html.
Jones, P.J., Levari, D., Bellet, B.W., McNally, R.J., 2021. Exposure to descriptions of traumatic events narrows the conceptual bracket of trauma. Journal of Experimental Psychology: Applied.
Jones, P.J., McNally, R.J., 2021. Does broadening one's concept of trauma undermine resilience? Psychol. Trauma Theo. Res. Pract. Policy. https://doi.org/10.1037/tra0001063. Advance online publication.
Kessler, R.C., Aguilar-Gaxiola, S., Alonso, J., Benjet, C., Bromet, E.J., Cardoso, G., Koenen, K.C., 2017. Trauma and PTSD in the WHO world mental health surveys. Eur. J. Psychotraumatol. 8, 1353383.
Lees, A.B., 2018. Yes, you can be traumatized by the media! Psychology Today. Retrieved from. https://www.psychologytoday.com/us/blog/surviving-thriving/201810/yes-you-can-be-traumatized-the-media.
Liu, H., Petukhova, M.V., Sampson, N.A., Aguilar-Gaxiola, S., Alonso, J., Andrade, L.H., World Health Organization World Mental Health Survey Collaborators, 2017. Association of DSM-IV posttraumatic stress disorder with traumatic experience type and history in the World Health Organization World Mental Health Surveys. JAMA Psychiatry 74 (3), 270–281.

McGrath, M.J., Randall-Dzerdz, K., Wheeler, M.A., Murphy, S., Haslam, N., 2019. Concept creepers: Individual differences in harm-related concepts and their correlates. Personal. Indiv. Diff. 147, 79–84.
McNally, R.J., 2003. Progress and controversy in the study of posttraumatic stress disorder. Annu. Rev. Psychol. 54 (1), 229–252.
McNally, R.J., 2009. Can we fix PTSD in DSM-V? Depress. Anxiety 26, 597–600.
McNally, R.J., 2015. Posttraumatic stress disorder and dissociative disorders. In: Blaney, P.H., Krueger, R.F., Millon, T. (Eds.), Oxford textbook of psychopathology. Oxford University Press, Oxford, England, pp. 191–221.
Pai, A., Suris, A.M., North, C.S., 2017. Posttraumatic stress disorder in the DSM-5: Controversy, change, and conceptual considerations. Behav. Sci. 7 (1), 7.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Robinson, J., Rosenzweig, C., Moss, A.J., Litman, L., 2019. Tapped out or barely tapped? Recommendations for how to harness the vast and largely unused potential of the Mechanical Turk participant pool. PLoS One 14 (12).
Shalev, A.Y., Gevonden, M., Ratanatharathorn, A., Laska, E., Van Der Mei, W.F., Qi, W., van Zuiden, M., 2019. Estimating the risk of PTSD in recent trauma survivors: results of the International Consortium to Predict PTSD (ICPP). World Psychiatry 18 (1), 77–87.
Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86 (2), 420.
Tolin, D.F., Foa, E.B., 2006. Sex differences in trauma and posttraumatic stress disorder: a quantitative review of 25 years of research. Psychol. Bull. 132, 959.
Vylomova, E., Murphy, S., Haslam, N., 2019. Evaluation of semantic change of harm-related concepts in psychology. In: Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, pp. 29–34.
Williams, M.T., 2015. The link between racism and PTSD. Psychology Today. Retrieved from. https://www.psychologytoday.com/us/blog/culturally-speaking/201509/the-link-between-racism-and-ptsd.