 doi:10.20944/preprints202010.0342.v1

Youth Exposure to Hate in the Online Space: An Exploratory Analysis

Nigel Harriman [1,*] , Neil Shortland [2] , Max Su [1] , Tyler Cote [3] , Marcia A. Testa [1,4] , Elena Savoia [1,4]

[1] Emergency Preparedness Research and Practice Program, Division of Policy Translation and Leadership Development Harvard T.H. Chan School of Public Health, Boston, MA
[2] Center for Terrorism and Security Studies University of Massachusetts Lowell, Lowell, MA
[3] Operation 250, Lowell, MA
[4] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA
* Correspondence: preparedness@hsph.harvard.edu

Abstract: Today's youth have almost universal access to the internet and frequently engage in social networking activities using various social media platforms and devices. This is a phenomenon that hate groups are exploiting when disseminating their propaganda. This study seeks to better understand youth exposure to hateful material in the online space by exploring predictors of such exposure including demographic characteristics (age, gender and race), academic performance, online behaviors, online disinhibition, risk perception, and parents/guardians' supervision of online activities. We implemented a cross-sectional study design, using a paper questionnaire, in two high schools in Massachusetts (USA), focusing on students 14 to 19 years old. Logistic regression models were used to study the association between independent variables (demographics, online behaviors, risk perception, parental supervision) and exposure to hate online. Results revealed an association between exposure to hate messages in the online space and time spent online, academic performance, communicating with a stranger on social media, and benign online disinhibition. In our sample, benign online disinhibition was also associated with students' risk of encountering someone online that tried to convince them of racist views. This study represents an important first step in understanding youth's risk factors of exposure to hateful material online.

Keywords: online hate; hate speech; online disinhibition; online safety

Introduction

Today's youth have almost universal access to the internet and frequently engage in social networking activities using various social media platforms and devices [1]. This is a phenomenon that hate groups are exploiting when disseminating their propaganda [2, 3]. Data gathered from a demographically balanced sample of over a thousand youth in the United States, showed that approximately half of them, within a study period of three months, experienced exposure to hateful material while online [4]. Analyses of interviews with right-wing extremists demonstrated that communication over the internet provides an effective networking method amongst their supporters. These groups use the internet to convey their racist messages and adopt communication strategies that are appealing to youth, with the scope of recruiting new members, including the use of images, videos, music, and online games [5, 6]. The online space also

 doi:10.20944/preprints202010.0342.v1

provides hate groups with the unique opportunity to portray an image of unity and identity [6, 7]. This online collective identity self-perpetuates as a welcoming space of like-minded individuals, providing sought-after validation for potential members' societal grievances and attracting socially isolated individuals to become members of a community that accepts them [8]. Furthermore, the anonymity of the internet creates an environment where hate groups speak or act out more radically compared to what they would do in person [9]. This is in part attributed to a psychological process referred to as online disinhibition, a process that most individuals experience when online compared to real life leading to lack of restraints and increased openness of expression [10, 11]. The literature describes two types of online disinhibition: toxic and benign disinhibition. Toxic disinhibition manifests as a propensity towards a variety of negative attitudes and behaviors such as anger, vicious criticism, outgroup hatred [12], cyberbullying [13-15], racism [16], and aggression [17, 18]. On the contrary, benign disinhibition is mostly described as a positive process by which individuals feel increased comfort in manifesting acts of kindness when online [19-21] compared to in person which, however, in some cases can lead to undesirable situations [13, 22]. There is concern that the internet may provide youth with a gateway to online hate communities and expose them to a dizzying array of sites containing hateful material [23]. As such, it is important to understand who are the most vulnerable to such exposure, so to equip them with the requisite knowledge to critically assess the material they may come across while online. Recent research has identified various psychological and behavioral factors that may put an individual at risk of exposure to hate online, such as: race, level of education, victimization, weak family attachment, low trust in government, and time spent on the internet [24-26]. This study seeks to contribute to this body of research focusing on the understanding of the predictors of youth's exposure to online hate by exploring the role of demographic characteristics, attitudes, risk perceptions and online behaviors.

Materials and Methods

We implemented a cross-sectional study design, using a paper questionnaire, in two high schools in Massachusetts (USA), gathering data from students 14 to 19 years old. The schools were participating in a training program focused on online safety and data for this study were derived from a convenience sample. The data presented in this manuscript refer to the baseline assessment conducted prior to the training, the analysis of the impact of the training will be object of a future publication. The questionnaire was implemented in December 2018 in one high school and in April 2019 in the other. A copy of the questionnaire can be found in Appendix A. Parents were provided with information on the study and opt-out forms one month prior to data collection. Consent to participate in the study was obtained from the students prior to responding to the questionnaire. The study protocol and the survey questions were deemed exempt by the Harvard T.H. Chan Institutional Review Board as well as by the ethical committees of the school districts where the study was implemented (approval number: IRB16-1757).

Independent variables

Demographic variables included: race, gender, and school year. Academic performance was measured by asking the respondent about their most frequent grade type (A, A-/B+, B, C or lower). Respondents were asked about the amount of time they spend using technology each day, which social media tool (i.e. YouTube, Snapchat, Instagram, etc.) they use and how frequently they use it, with answer options ranging from 1 (never) to 6 (all the time). Overall social media

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2020 doi:10.20944/preprints202010.0342.v1

use was measured by a summative score created by converting the distribution of frequency of use for each social media platform into a Normal Standardized Distribution and adding the resultant scores across all social media platforms. Respondents were asked how many of their social media followers they knew in person, and if they had recently removed any strangers from such followers; similar questions were asked about respondents' friends' social media behaviors, under the assumption that they would be less likely to misreport friends' habits compared to their own. Online disinhibition was measured using the Online Disinhibition Scale after adapting the questions to the young age of the study population [13]. The factor structure of the questions was assessed with a factor analysis using principal component analysis for factor extraction, and as a result a scale with a score ranging from 7 to 28 was formed, with higher values indicating a more disinhibited behavior. Perception of online risk was measured by asking the respondent to rate the risk of seven online scenarios. In this case as well, a factor analysis was performed to assess the structure of the questions, and as a result a scale was created with score ranging from 7 to 35, with lower values indicating lower risk perception. For both scales, Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy and Bartlett's test of sphericity were used to test for the suitability of the data for factor analysis. Cronbach alpha was computed to assess the scales' reliability. Questions about engagement in risky online behaviors were also asked, such as chatting with strangers or sharing personal information with online contacts. Finally, respondents were asked about parents' supervision of their online activities and if they had a trusted adult to ask for help in the case they encountered an uncomfortable situation online.

Dependent variables

Two dependent variables were created to measure exposure to online hate during the two months prior to the survey. Respondents were asked to report how frequently they had come across insulting verbal or written expressions against a specific group because of their race, religion, disability, sexual orientation, ethnicity, gender, or gender identity (*exposure to online hate messages*). Respondents were also asked if, in the same time period, they encountered *someone trying to convince them of racist views*. Both variables were dichotomized as yes/no.

Statistical analyses

Simple and multiple logistic regression models were used to study the association between the independent and dependent variables. A Box-Tidwell procedure was used to confirm that the benign disinhibition score, the only continuous variable in the model, had a linear relationship to the log odds of each dependent variable. Independent variables were included in the multiple model when a statistically significant association was found in the simple models (p-value < 0.05). Gender, race, and school year were included regardless of their significance in the final models because of theoretical relevance. Hosmer-Lemeshow tests [27] were used to assess the goodness of fit of the models. Race and gender were also tested as interaction terms. Data analysis was performed using Stata Statistical Software: Release 15.1. College Station, TX: StataCorp LLC.

Results

We provide below the sample characteristics and descriptive statistics for social media use and online behaviors, adult supervision of online activities, and exposure to hate online. Following, we present the results of the factor analyses for the online disinhibition and risk

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2020 doi:10.20944/preprints202010.0342.v1

perception scales, and the results of the simple and multiple logistic regression models for exposure to online hate.

Sample characteristics and descriptive statistics

We gathered data from a convenience sample of 320 individuals, the majority of which were female (58%). Major categories of race were distributed as follows: white (35%), mixed race (21%), and Hispanic (18%). The sample primarily included students from grades 9 (27%) and 10 (66%), 4% were in 11th grade, and 3% in 12th grade. The majority of the respondents (75%) reported spending over three hours a day interacting with technology, excluding schoolwork. YouTube, Snapchat, and Instagram were the most frequently used social media tools with over 90% of respondents reporting their use. Only 7% of the respondents knew all of their social media followers, the majority of them (63%) reported that, in the two months prior to the survey, they chatted with someone on social media that they had not met in person and 45% reported they believe their friends did so as well. During the same timeframe, 67% removed a follower from their social media account they had not met in person, and 48% reported to have shared personal information, such as their school or town name, when posting on social media. Among those playing video games online, 52% reported that in the two months prior to the survey they chatted with someone they did not know while gaming. Regarding parents' supervision, the majority (57%) of students had parents that occasionally asked about their online activities, but only 25% reported that their parents had rules for what they did online and checked on them to make sure they followed the rules. The majority (64%) of students reported they had a trusted adult they could ask for help if they experienced an online situation that made them feel uncomfortable. Fifty-seven percent of students reported to have come across hate messages on social media or on a website in the two months prior to the survey and 12% reported to have encountered someone online that tried to convince them of racist views during the same time period. More details on the descriptive statistics are provided in Table 1.

Table 1: Sample characteristics

School year (Q1)	**N (%)**	**Shared personal information online (Q11)**	**N (%)**
9th Grade	86 (27%)	I don't have a social media account	7 (2%)
10th Grade	211 (66%)	Never	161 (50%)
11th Grade	14 (4%)	Sometimes	114 (36%)
12th Grade	9 (3%)	Often	38 (12%)
Gender (Q4)	**N (%)**	**Removed strangers from social media followers (Q9)**	**N (%)**
Male	133 (42%)	No	105 (33%)
Female	181 (58%)	Yes	215 (67%)
Academic performance (Q2)	**N (%)**	**Parents ask about online activities (Q15)**	**N (%)**
A	51 (16%)	Parents do not ask	71 (22%)
A-/B+	115 (36%)	Parents occasionally ask	181 (57%)
B	48 (15%)	Parents frequently ask	49 (15%)
B-/C+	70 (22%)	I think my parents check on my devices	12 (4%)
C or Lower	36 (11%)	I am never online	6 (2%)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2020 doi:10.20944/preprints202010.0342.v1

Race (Q5)	**N (%)**	**Friends communicated with strangers (Q12)**	**N (%)**
White	108 (35%)	I am not sure	78 (24%)
Asian	22 (7%)	Never	24 (8%)
Black	26 (8%)	Sometimes	143 (45%)
Cape Verdean	16 (5%)	Often	75 (23%)
Hispanic	56 (18%)		
Haitian	7 (2%)	**Play video games online (Q13)**	**N (%)**
Mixed Race	68 (21%)	Never	141 (44%)
Don't Know / Rather not Say	7 (2%)	Sometimes	95 (30%)
Other	10 (3%)	Often	84 (26%)
Time spent online (Q6)	**N (%)**	**Chatted with strangers while gaming (Q14)**	**N (%)**
Less than 1 hour	7 (2%)	I do not play video games	71 (22%)
More than 1 but less than 3 hours	74 (23%)	Never	120 (38%)
3 to 6 hours	155 (49%)	Sometimes	76 (24%)
More than 6 hours	83 (26%)	Often	53 (17%)
Social media followers met in person (Q8)	**N (%)**	**Parental rules for online activity (Q16)**	**N (%)**
I am not sure how many	35 (11%)	Do not have rules	130 (41%)
Some of them	67 (21%)	Have a few rules but don't check to see if they are followed	111 (35%)
Most of them	188 (60%)	Have a few rules and check	67 (21%)
All of them	22 (7%)	Have many rules and check	11 (3%)
Communicated with strangers on social media (Q10)	**N (%)**	**Trusted adult (Q18)**	**N (%)**
I don't have a social media account	6 (2%)	No	34 (11%)
Never	111 (35%)	Yes	203 (63%)
Sometimes	156 (49%)	Not sure	56 (18%)
Often	46 (14%)	It depends on the situation	27 (8%)
Exposed to hate messages online (Q22 – social media or website)	**N (%)**	**Encountered someone trying to convince the respondent of racist views (Q17)**	**N (%)**
Yes	182 (57%)	Yes	38 (12%)
No	138 (43%)	No	282 (88%)
Risk Perception Scale (Q19)		**Benign Online Disinhibition Scale (Q7a, b, d, e, g, h, k)**	
Mean	29.35	Mean	18.44
Standard Deviation	4.99	Standard Deviation	4.72
Median	31	Median	18
Range	7-35	Range	7-28

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2020 doi:10.20944/preprints202010.0342.v1

Factor analyses and descriptive statistics of risk perception and online disinhibition

Risk perception scale

Bartlett's Test of Sphericity (χ^2=865.397, df =21, p<0.01) and Kaiser-Meyer-Olkin measure of sampling adequacy (0.834) indicated that the data on the seven questions designed to measure risk perception were suitable for factor analysis. A factor analysis was computed resulting in one factor with an eigenvalue greater than 1, all items had a factor loading greater than 0.4 and 52% of the variance in the data was explained by the model. Cronbach's alpha was 0.84. Risk perception scores were negatively skewed with a mean of 29.3 (SD=5) and a median of 31 (range: 7-35). In the simple and multiple models, risk perception was examined as a dichotomous variable. Individuals with scores less than or equal to 28 (25th percentile) were defined as having "low risk perception" (n=103). Individuals with scores greater than 28 were defined as having "high risk perception" (n=217).

Benign disinhibition scale

Bartlett's Test of Sphericity (χ^2= 553.869, df=55, p<0.01) and Kaiser-Meyer-Olkin measure of sampling adequacy (0.8) indicated that the data on the online disinhibition questions were suitable for factor analysis. A factor analysis was computed on the 11 questions resulting in two factors with eigenvalues greater than 1 which explained 45% of the variance in the data. After oblique promax rotation, one question with a factor loading below 0.4 (q7f – see Supplementary Questionnaire) was discarded. The resulting two subscales were similarly structured to the scale developed by Udris et. al. [13] The benign disinhibition subscale contained 7 questions (7a, b, d, e, g, h, k – see Supplementary Questionnaire) and had a Cronbach's alpha of 0.72. The toxic disinhibition subscale contained 3 questions. Due to its low internal consistency (Cronbach's alpha = 0.53), toxic disinhibition was not included in the simple and multiple analysis. Benign online disinhibition scores had a mean of 18.4 (SD= 4.7) and a median of 18 (range: 7 to 28). Box-Tidwell test results from the simple regression analyses supported assumption that the distribution of benign online disinhibition scores were linear to the log odds of respondents' exposure to hate messages online (p=0.4) and to the experience of encountering someone online that tried to convince them of racist views (p=0.8).

Simple models

Exposure to hate messages online

In the simple regression models, the following variables were significantly associated with the dependent variable - *exposure to hate messages online:* time spent online (OR=2.3, 95% CI 1.4-3.8), removing a follower from a social media account - whom the respondent had not met in person (OR=1.9, 95% CI 1.2-3), communicating with someone on social media that the respondent had not met in person (OR=2.1, 95% CI 1.3-3.3), presence of parental rules for online activities (OR=1.6, 95% CI 1-2.5), benign online disinhibition (OR=1.09, 95% CI 1.04-1.14), and good academic performance (OR=2.3, 95% CI 1.1-4.6). Detailed results for the simple models of exposure to hate messages online can be found in Table 2.

Encountering someone trying to convince the respondent of racist views

In the simple regression models, the following variables were significantly associated with the dependent variable - *encountering someone online that tried to convince the respondent of*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2020 doi:10.20944/preprints202010.0342.v1

racist views: school year (OR=0.5, 95% CI 0.2-0.9), playing video games online (OR=2.1, 95% CI 1-4.4), and benign online disinhibition (OR=1.19, 95% CI 1.09-1.29). The only significant categorical predictor with more than two categories was chatting - while gaming online - with someone the respondent had never met in person (p=0.041). Detailed results for the simple models of exposure to hate messages online can be found in Table 2.

Table 2. Simple models

Independent variables	Exposed to hate messages online	Encountered someone trying to convince the respondent of racist views
Dichotomous and Continuous Predictors	**OR (95% CI)**	
School year (10th grade and above vs. 9th grade)	0.8 (0.5-1.3)	0.5 (0.2-0.9) *
Race (White vs. Non-white)	1 (0.6-1.7)	1.5 (0.8-3.1)
Gender (Female vs. Male)	1.5 (0.9-2.3)	0.7 (0.3-1.3)
Time spent online (≥3 hours vs. < 3 hours)	2.3 (1.4-3.8) **	1.6 (0.7-3.8)
Removed a stranger from your social media followers (Yes vs. No)	1.9 (1.2-3) *	2 (0.9-4.5)
Communicated with strangers on social media (Sometimes/ Often vs. Never/No social media account)	2.1 (1.3-3.3) **	2 (0.9-4.4)
Shared personal information online (Sometimes/ Often vs. Never/ No social media account)	1.3 (0.8-2)	1.4 (0.7-2.8)
Played online video games (Sometimes/ Often vs. Never)	1.2 (0.8-1.8)	2.1 (1-4.4) *
Parents have rules for what you do online (Any level of rule vs. Do not have rules)	1.6 (1-2.5) *	0.5 (0.3-1)
Parents ask what you do online (Any frequency of asking vs. Do not ask)	1.5 (0.9-2.6)	1.7 (0.7-4.1)
Benign Online Disinhibition	1.09 (1.04-1.14) **	1.19 (1.09-1.29) **
Risk perception (≤25th Percentile vs. > 25th Percentile)	1 (0.6-1.7)	1.1 (0.5-2.3)
Academic performance (≥ C+ vs. ≤ C)	2.3 (1.1-4.6) *	1.1 (0.4-3.3)
Followers met in person (Not sure/Some of them vs. Most of them/ All of them)	1.3 (0.8-2)	1.2 (0.6-2.6)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2020 doi:10.20944/preprints202010.0342.v1

Categorical Predictors with 3+ Categories	Chi-Squared (df); p-value	
Friends communicate with someone on social media they had not met in person	6.92 (df=3); p-value=0.075	5.69 (3); p-value=0.128
Chatting with someone you have never met while gaming online	0.14 (df=2); p-value=0.933	6.39 (2); p-value=0.041
Having a trusted adult to speak to in case you come across something unsafe online	0.37 (df=3); p-value=0.946	4.89 (3); p-value=0.180
Social media use quartiles	1.07 (df=3); p-value=0.785	1.06 (3); p-value=0.786

*$p < 0.05$; **$p < 0.01$

Multiple models

Exposure to hate messages online

The overall LR chi-square test statistic for the multiple model exploring the association between the independent variables and exposure to hate messages online was significant (χ^2=40.54, df=9, p<0.01). Hosmer-Lemeshow Goodness of Fit test results confirmed that the model was a good fit for the data (χ^2=6.62, df=8, p=0.578). Time spent online was associated with increased odds of exposure to online hate messages - youth that spent three or more hours a day online had 2.4 times the odds of reporting exposure to hate messages (seen either on a website or social media), (OR=2.4, 95% CI 1.3-4.3) compared to those who spent less than 3 hours a day online. The odds of reporting exposure to hate messages among those who communicated with someone on social media that they had not met in person were 1.7 times that of those who had not done so (OR=1.7, 95% CI 1-2.9). Benign online disinhibition was associated with reporting exposure to hate messages online – each one-unit increase in score on the benign disinhibition scale resulted in a 6% increase in the odds (OR=1.06, 95% CI 1-1.12) of reporting exposure to such messages. Good academic performance was also associated with exposure to online hate messages, students who reported receiving grades greater than a C had 3.4 times the odds of reporting exposure to such messages compared to students who were receiving Cs or lower grades (OR=3.4, 95% CI 1.5-7.7). Gender and race were used to study their interaction with the following variables: time spent online, benign disinhibition and communicating with a person not met in person while online. None of these interactions resulted to be significant. Detailed results for the multiple models of exposure to online hate messages can be found in Table 3.

Encountering someone trying to convince the respondent of racist views

The overall LR chi-square test statistic for the model investigating the association between independent variables and encountering someone trying to convince the respondent of racist views was significant (χ^2=26.36, df=7, p<0.01). Hosmer-Lemeshow Goodness of Fit test results confirmed that the model was an appropriate fit for the data (χ^2=6.13, df=8, p=0.6325). Only benign online disinhibition was associated with students' risk of encountering someone trying to convince the respondent of racist views. Each one unit increase in benign disinhibition score resulted in a 19% increase in the odds (OR=1.19, 95% CI 1.09-1.31) of experiencing this

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2020 doi:10.20944/preprints202010.0342.v1

situation. There were no significant interaction terms observed. Detailed results can be found in Table 3.

Table 3. Multiple models

Independent variables	Exposed to hate messages online	Encountered someone trying to convince the respondent of racist views
Dichotomous and Continuous Predictors	**OR (95% CI)**	
School year (10th grade and above vs. 9th grade)	0.8 (0.5-1.5)	0.6 (0.3-1.4)
Race (White vs. Non-white)	1.1 (0.6-1.8)	1.9 (0.9-4)
Gender (Female vs. Male)	1.5 (0.9-2.5)	1 (0.4-2.7)
Time spent online (≥3 hours vs. < 3 hours)	2.4 (1.3-4.3) **	NA
Removed a stranger from your social media followers (Yes vs. No)	1.6 (0.9-2.7)	NA
Communicated with strangers on social media (Sometimes/ Often vs. Never/No social media account)	1.7 (1-2.9) *	NA
Played online video games (Sometimes/ Often vs. Never)	NA	1.2 (0.4-3.8)
Parents have rules for what you do online (Any level of rule vs. Do not have rules)	1.4 (0.8-2.3)	NA
Benign Online Disinhibition	1.06 (1-1.12) *	1.19 (1.09-1.31) **
Academic performance (≥ C+ vs. ≤ C)	3.4 (1.5-7.7) **	NA
Categorical Predictors with 3+ Categories	**Chi-Squared (df); p-value**	
Chatting with someone you have never met while gaming online	NA	0.52 (2); p-value=0.77

*$p < 0.05$; **$p < 0.01$*

Discussion

Technology has become increasingly important in the lives of adolescents who are heavy users of various forms of electronic communication such as instant messaging, e-mail, social media, and sites where they share opinions, photos, and videos [28]. Although teens usually find valuable educational support and information on the internet, they can also be exposed to online propaganda, hate messages, racism and negative influences. A US-based national study demonstrated that amount of general technology use and age are predictive factors for almost all technology-based violent experiences and exposures [29]. The presence, form, and function of extremist material available on the internet has been extensively discussed by academics and practitioners [30-34]. Some have theorized that access to such material influences the likelihood that an individual will, eventually, engage in hateful or violent behavior [29, 35, 36]. However, there is limited knowledge on the risk factors that lead an individual to be exposed to online hate in the first place and on the consequences of such exposure. A better understanding of the risk

 doi:10.20944/preprints202010.0342.v1

factors could be useful to develop educational interventions and prevention programs that equip youth with the knowledge they need to appropriately react to such material.

In our study, we found that the more time youth spent online the more likely they were to be exposed to hate in the online space. This result is consistent with previous literature [24-26]. Not surprisingly, communicating with strangers online was associated with increased risk of being exposed to hate. Interestingly, good academic performance was also associated with increased risk, this may be due to increased awareness and ability to recognize the online material as hateful or by some interest on the topic expressed by higher educated youth, as found in previous research [25]. Finally, our data indicate that the more individuals felt disinhibited online, *"loosening up, feeling less restrained, and expressing themselves more openly,"* the more likely they were to be exposed to hateful propaganda and to encounter individuals attempting to convince them of racist views. Our results raise interesting questions about the nature of disinhibition and the underlying processes that moderate the relationship between disinhibition and exposure to hate. What is of specific interest here, however, is that benign disinhibition was significantly associated with exposure to hateful material. We believe that while toxic disinhibition may be associated with active engagement in hate and harmful activities, benign disinhibition as a whole may be associated with passive exposure to hate [12]. These preliminary results pose intriguing and critical questions for future research. First, does someone exhibiting *only* benign disinhibited behaviors stray free from the innate risks of hateful content? Second, what role does mere exposure play on one's own future behavior, feelings, and psychology?

Past research has focused on long-term exposure to hateful content online showing that it might reinforce discriminatory views and could lead to developing defensive and hyper-vigilant attitudes [26, 36, 37]. In many prevention spheres, prevention efforts have focused on educating about what risks are present online, and indeed the prevalence and nature of nefarious online actors and groups. Our results pose important implications for efforts to increase online safety in that educational initiatives need to prioritize self-reflection and self-awareness as much as content-based knowledge. Preliminary studies into online safety education programming show that initiatives focused only on enhancing knowledge might be missing the spot for long-term education [38]. In an evaluation of existing internet safety resources and programs, results have shown that experiencing a risky online situation isn't about "lack of knowledge," but rather omission of the necessary skills needed when navigating the internet [39]. Such skills may include learning how to limit one's own time online, self-awareness of disinhibited behavior, and avoiding risky situations, such as sharing personal information online or engaging with strangers.

When reflecting on the findings of this study, it is imperative to recognize its limitations. The primary limitation is its cross-sectional design. As such, the observed relationship between independent and dependent variables, albeit plausible, should not be assumed as causal. Additionally, due to the cross-sectional nature of these data, the direction of the observed relationship between exposure and outcome can only be speculated. As some of the exposures and outcomes could be perceived as negative, there is potential for social desirability bias to influence students' responses. As a non-random convenience sample, it is important to acknowledge that these results are not necessarily generalizable outside of the sample, and that selection bias may have also influenced the results. Finally, there may be limitations with the measurement of online disinhibition within our sample due to the lack of variance explained by the scale and the fact that toxic disinhibition could not be reliably measured. Yet, we believe our findings generate important preliminary recommendations for the development of educational activities aimed at improving online safety that should focus on teaching youth how to limit the

time they spend online, helping them recognize the disinhibition effect of the internet and how passive online behaviors may also generate risk and influence decisions they make when online.

Conclusions

Our study of a population of high school students found an association between exposure to hate messages in the online space and time spent online, academic performance, communicating with a stranger on social media, and benign online disinhibition. In our sample, benign online disinhibition was also associated with students' risk of encountering someone online that tried to convince them of racist views. This study represents an important first step in understanding youth's risk factors of exposure to hateful material online.

Author Contributions: N.H. Performed the statistical analyses and drafted the introduction, methods, results, and tables. N.S. and T.C. drafted the discussion and conclusions. M.S. Supervised the data analysis and provided feedback for the results and methods sections. M.A.T. Developed the study design and supervised the data analysis E.S. Developed the study design and provided feedback on all sections of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: Research reported in this publication was supported by *the National Institute of Justice* under award number 2016-ZA-BX-K001 Evaluation of the Peer to Peer (P2P): Challenging Extremism Initiative and award number 2018-2A-CX-0002 Operation250: An Evaluation of a Primary Prevention Campaign focused on Online Safety and Risk Assessment. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Justice.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Leggett, S., Childhood 2017: For children and teenagers, it is increasingly all about mobile. Childwise: 2017.
2. Corb, A.; Grozelle, R. S., A New Kind of Terror: Radicalizing Youth in Canada. *Journal EXIT-Deutschland* **2014,** *1* (2014).
3. Mughal, S. Radicalisation of young people through social media 2016. https://www.internetmatters.org/hub/expert-opinion/radicalisation-of-young-people-through-social-media/.
4. Hawdon, J.; Oksanen, A.; Räsänen, P., Chapter 9 - Victims of Online Groups: American Youth's Exposure to Online Hate Speech. In *The Causes and Consequences of Group Violence: From Bullies to Terrorists*, Hawdon, J.; Ryan, J.; Lucht, M., Eds. Lexington Books: London, UK, 2014; pp 165–182.
5. Bliuc, A.-M.; Faulkner, N.; Jakubowicz, A.; McGarty, C., Online networks of racial hate: A systematic review of 10 years of research on cyber-racism. *Computers in Human Behavior* **2018,** *87*, 75-86.
6. Gerstenfeld, P. B.; Grant, D. R.; Chiang, C.-P., Hate online: A content analysis of extremist Internet sites. *Analyses of Social Issues and Public Policy (ASAP)* **2003,** *3* (1), 29-44.
7. Ducol, B., Uncovering the French-speaking jihadisphere: An exploratory analysis. *Media, War & Conflict* **2012,** *5* (1), 51-70.

8. Davies, G.; Neudecker, C.; Ouellet, M., Toward A Framework Understanding of Online Programs For Countering Violent Extremism. *Journal for Deradicalization* **2016,** *Spring* (6), 51-86.
9. Koehler, D., The Radical Online: Individual Radicalization Processes and the Role of The Internet. *Journal for Deradicalization* **2014,** *1* (2014), 116-134.
10. Suler, J., The Online Disinhibition Effect. *Cyberpsychology & Behavior* **2004,** *7* (3), 321-326.
11. Joinson, A., Causes and implications of disinhibited behavior on the Internet. In *Psychology and the Internet: Intrapersonal, interpersonal, and transpersonal implications.*, Gackenbach, J., Ed. Academic Press: San Diego, CA, US, 1998; pp 43-60.
12. Wachs, S.; Wright, M.F. Associations between Bystanders and Perpetrators of Online Hate: The Moderating Role of Toxic Online Disinhibition. *Int. J. Environ. Res. Public Health* **2018**, *15*, 2030.
13. Udris, R., Cyberbullying among high school students in Japan: Development and validation of the Online Disinhibition Scale. *Computers in Human Behavior* **2014,** *41*, 253-261.
14. Wright, M. F.; Harper, B. D.; Wachs, S., The associations between cyberbullying and callous-unemotional traits among adolescents: The moderating effect of online disinhibition. *Personality and Individual Differences* **2019,** *140*, 41-45.
15. Barlett, C. P.; Helmstetter, K. M., Longitudinal relations between early online disinhibition and anonymity perceptions on later cyberbullying perpetration: A theoretical test on youth. *Psychology of Popular Media Culture* **2018,** *7* (4), 561-571.
16. Keum, B. T.; Miller, M. J., Racism on the Internet: Conceptualization and recommendations for research. *Psychology of Violence* **2018,** *8* (6), 782-791.
17. Kurek, A.; Jose, P. E.; Stuart, J., 'I did it for the LULZ': How the dark personality predicts online disinhibition and aggressive online behavior in adolescence. *Computers in Human Behavior* **2019,** *98*, 31-40.
18. Lapidot-Lefler, N.; Barak, A., Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior* **2012,** *28* (2), 434-443.
19. Lapidot-Lefler, N.; Barak, A., The benign online disinhibition effect: Could situational factors induce self-disclosure and prosocial behaviors? *Cyberpsychology* **2015,** *9* (2).
20. Bareket-Bojmel, L.; Shahar, G., Emotional and Interpersonal Consequences of Self-Disclosure in a Lived, Online Interaction. *Journal of Social and Clinical Psychology* **2011,** *30* (7), 732-759.
21. Amichai-Hamburger, Y.; Kingsbury, M.; Schneider, B. H., Friendship: An old concept with a new meaning? *Computers in Human Behavior* **2013,** *29* (1), 33-39.
22. Charaschanya, A.; Blauw, J., A Study of The Direct and Indirect Relationships between Online Disinhibition and Depression and Stress Being Mediated by The Frequency of Cyberbullying from Victim and Perpetrator Perspectives. *Scholar : Human Sciences* **2018,** *9* (2), 275-301.
23. Livingstone, S.; Haddon, L.; Görzig, A.; Ólafsson, K. *Risks and safety on the internet: the perspective of European children: full findings and policy implications from the EU Kids Online survey of 9-16 year olds and their parents in 25 countries*; ISSN 2045-256X; LSE: London, 2011.
24. Costello, M.; Barrett-Fox, R.; Bernatzky, C.; Hawdon, J.; Mendes, K., Predictors of Viewing Online Extremism Among America's Youth. *Youth & Society* **2018,** *52* (5), 0044118X1876811-727.

25. Costello, M.; Hawdon, J.; Ratliff, T.; Grantham, T., Who views online extremism? Individual attributes leading to exposure. *Computers in Human Behavior* **2016,** *63*, 311-320.
26. Oksanen, A.; Hawdon, J.; Holkeri, E.; Näsi, M.; Räsänen, P., Exposure to Online Hate among Young Social Media Users. In *Soul of Society: A Focus on the Lives of Children & Youth*, Warehime, M. N., Ed. Emerald Group Publishing Limited: 2014; Vol. 18, pp 253-273.
27. Hosmer, D. W.; Lemeshow, S.; Sturdivant, R. X., *Applied Logistic Regression, third edition*. 3rd ed. ed.; John Wiley and Sons: Hoboken, NJ, 2013.
28. Kaveri, S.; Patricia, G., Online Communication and Adolescent Relationships. *The Future of Children* **2008,** *18* (1), 119-146.
29. Ybarra, M. L.; Mitchell, K. J.; Korchmaros, J. D., National trends in exposure to and experiences of violence on the Internet among children. *Pediatrics* **2011,** *128* (6), e1376-e1386.
30. Bowman-Grieve, L., Exploring "Stormfront": A Virtual Community of the Radical Right. *Studies in Conflict and Terrorism* **2009,** *32* (11), 989-1007.
31. Bowman-Grieve, L.; Conway, M., Exploring the form and function of dissident Irish Republican online discourses. *Media, War & Conflict* **2012,** *5* (1), 71-85.
32. Weimann, G. *Www.terror.net: How Modern Terrorism Uses the Internet*; Special Report 116; United States Institute of Peace: 2004.
33. Weimann, G., *Terror on the Internet: The New Arena, the New Challenges*. 1 ed.; United States Institute of Peace Press: Washington DC, US, 2006.
34. Weimann, G., Cyber- Fatwas and Terrorism. *Studies in Conflict and Terrorism* **2011,** *34* (10), 765-781.
35. Baaken, T.; Schlegel, L., Fishermen or Swarm Dynamics? Should we Understand Jihadist Online-Radicalization as a Top-Down or Bottom-Up Process? *Journal for Deradicalization* **2017,** *Winter* (13), 178-212.
36. Wachs, S.; Wright, M.F.; Sittichai, R.; Singh, R.; Biswal, R.; Kim, E.-M.; Yang, S.; Gámez-Guadix, M.; Almendros, C.; Flora, K.; Daskalou, V.; Maziridou, E. Associations between Witnessing and Perpetrating Online Hate in Eight Countries: The Buffering Effects of Problem-Focused Coping. *Int. J. Environ. Res. Public Health* **2019**, *16*, 3992.
37. Lee, E.; Leets, L., Persuasive Storytelling by Hate Groups Online: Examining Its Effects on Adolescents. *American Behavioral Scientist* **2002,** *45* (6), 927-957.
38. Foxman, A. H., *Viral hate : containing its spread on the Internet*. Palgrave Macmillan: New York, NY, 2013.
39. Jones, L. M.; Mitchell, K. J.; Walsh, W. A. *Evaluation of Internet Child Safety Materials Used by ICAC Task Forces in School and Community Settings, Final Report*; 242016; Crimes Against Children Research Center (CCRC) - University of New Hampshire: US Department of Justice, 2013.