Characteristics of People Who Engage in Online Harassing Behavior

Song Mi Lee
songmil@umich.edu
School of Information, University of Michigan
Ann Arbor, Michigan, USA

Cliff Lampe
cacl@umich.edu
School of Information, University of Michigan
Ann Arbor, Michigan, USA

J.J. Prescott
jprescott@umich.edu
Law School, University of Michigan
Ann Arbor, Michigan, USA

Sarita Schoenebeck
yardi@umich.edu
School of Information, University of Michigan
Ann Arbor, Michigan, USA

ABSTRACT

Conflict in online spaces can often lead to behaviors that may be categorized as "harassment." We asked 307 U.S. adults to self-report if they have ever engaged in aggressive online conflict. Using logistic regressions, we examine what psychosocial characteristics predict which users would report engaging in behaviors that are commonly labeled as "harassment." We find that psychological factors such as impulsivity, reactive aggression, and premeditated aggression distinguish those who never thought of, those who only imagined, and those who carried out harassing behavior. Demographic factors other than age do not have significance, contrary to the results of prior studies. Design interventions that address propensities to perpetrate harassment might reduce harm but also raise ethical and moral concerns about the nature of harassment and the disposition toward it.

CCS CONCEPTS

• **Human-centered computing** → Empirical studies in collaborative and social computing.

KEYWORDS

online harassment, online harasser, online conflict, aggression

ACM Reference Format:
Song Mi Lee, Cliff Lampe, J.J. Prescott, and Sarita Schoenebeck. 2022. Characteristics of People Who Engage in Online Harassing Behavior. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '22 Extended Abstracts), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3491101.3519812

1 INTRODUCTION

Theories of crime and deviance propose that understanding a potential perpetrator's psychological traits, motives, and needs can effectively reduce the likelihood of an offense [3]. Likewise, understanding the psychosocial characteristics of people who engage in activities commonly bundled under the label "online harassment" may be fruitful for addressing and remediating those behaviors [6]. Exploring interventions to reduce harassing behaviors can lessen the reliance on content moderation, lower the burden on targets of harassment post hoc, and also protect perpetrators from engaging in behaviors they might regret later and from becoming the target of backlash or retributive harassment. Toward this aim, we ask: what are the characteristics of adults who self-report engaging in online behavior that is often defined as "harassing?"

Recruiting adults who have been perpetrators of online harassment has been a major challenge in this line of research due to definitional confusion and the social stigma surrounding harassment. Consequently, this group has been underexamined—existing work is limited to those on specific online platforms (e.g., [25]) or based on speculation emerging from either targets' accounts (e.g., [20, 33]) or research on adolescent cyberbullying [10]. One recent study of "trolls" indicates that harassing behavior is more widespread than might have been thought in earlier narratives [7].

Focusing on behaviors rather than labels provides an alternative approach to characterizing people who engage in harassing behavior among everyday Internet users. Online harassment, in essence, is an interpersonal or group conflict behavior resulting from incompatible or opposing interests, aims, or needs [29]. We use this conflict framing in our survey language instead of directly asking participants to self-report "online harassment" or activities that can be categorized as such. We conducted an online survey in which we asked 307 U.S. adults if they had engaged in any conflict with someone on the Internet either to show their disapproval, to upset the other person, or in anger. We find that psychological factors predict self-reported perpetration better than demographic factors. We identify significant psychological differences among those who never considered engaging in aggressive online conflict, those who were tempted to but did not, and those who did. We discuss opportunities for designing platforms that better adapt and support varied needs and goals among users who engage in conflict, and we then reflect on the ethical implications of considering psychosocial characteristics in online harassment interventions.

CHI '22 Extended Abstracts, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9156-6/22/04…$15.00
https://doi.org/10.1145/3491101.3519812

2 BACKGROUND

2.1 Addressing Challenges in Self-Reports of Harassment Perpetration

Not only researchers but also perpetrators, targets, and bystanders disagree about the definition of online harassment. In particular, individuals accused of being harassers often do not agree with labeling their behavior as harassment [13, 15, 20]. For instance, blocked Twitter users report many cases where they think disagreements were unfairly portrayed as harassment by others [15]. Some more consciously see harassment as a justified act of moral outrage, blaming the target's wrongdoing [20]. Similar confusions and denials are also common among harassment targets and bystanders. According to a recent Pew report, 36% of targets would not personally call their experience online harassment, and 21% report being unsure. Men are much less likely to describe their negative experiences as harassment than women [33]. Adolescents and parents eschew the labels of harassment or bullying, instead framing candidate incidents as unfortunate but just common "drama" [19, 23]. Such contested perceptions can lead to inconsistent operationalization and unreliable self-report analyses.

It is especially challenging to capture and interpret adult self-reports of engaging in harassment because of the social stigma surrounding it. Recent studies suggest that anyone can engage in harassing behavior. For instance, Cheng et al. [7] show that ordinary Internet users regularly make trolling comments when in a negative mood. Song et al. [26] observe that offensive speech can be contagious when individuals utter swear words to mimic peers. However, media and scholarship predominantly depict online harassers as "trolls" or individuals with inherently juvenile, narcissistic, sadistic, psychopathic, or otherwise socially maladjusted dispositions [8, 17]. Because adults are considered much more likely to give socially desirable responses than younger populations [28], it is hard to expect that adults would willingly and honestly admit to having engaged in socially unacceptable behaviors like harassment [20].

Acknowledging that anyone may be at risk of engaging in harassment, and thus not dismissing harassment as fringe behavior, we frame online harassment as interpersonal conflict in our survey. A few studies suggest that interpersonal conflict may be characteristic of incivility or harassment [1, 14, 20]. Allen [1], by analyzing different cyberbullying schemas in adolescents and adults, argues that conflict, aggression, and bullying are fluid, dynamic, nuanced, and overlapping phenomena. She concludes that "the focus on defining a construct that is subjectively understood may not be as helpful as studying a variety of overlapping phenomena" (p. 177) and that context and participant perspectives should play a role in solving the bullying problem. Thus, using the overarching, more lenient, and less contestable notion of interpersonal conflict could reduce confusion, resistance, or social stigma in questionnaire-based research.

2.2 Personal Factors of Aggression

As online harassment can range from a brief, single incident between strangers to sustained threats, it can present in the world as an angry but regretful outburst, a calculated operation, a lark for "the lulz," or anything in between. Traditional aggression research outside the online context has organized various personal factors applicable to the general (i.e., not necessarily clinical or antisocial) population to account for the pervasiveness and heterogeneity of aggression [2]. The most discussed factors include the following:

- Impulsivity: a predisposition toward rapid, unplanned reactions to internal or external stimuli without regard to the negative consequences of these reactions [27]
- Moral disengagement: a cognitive process individuals use to justify behaviors they know are wrong [11, 12]
- Disinhibition: a tendency to act in an under-controlled manner [22]
- Reactive/proactive aggression: Reactive aggression indicates an angry defensive response to a perceived threat, fear, or provocation. Proactive aggression involves more hostile and goal-directed components with manipulative, callous, or cold-blooded acts to attain a goal [4].
- Impulsive/premeditated aggression: Impulsive aggression refers to uncontrolled, emotionally charged aggressive acts which are spontaneous, either provoked or out of proportion to the provocation. It often accompanies feelings of regret [5]. Premeditated aggression concerns planned or conscious aggressive acts, not spontaneous or related to an agitated state [4].
- Demographics: gender, age, socio-economic status, etc. [16]

While research on demographic factors tends to show mixed or statistically insignificant relationships to measures of aggression, studies confirm the importance of psychosocial factors in understanding aggression. In particular, reactive/proactive and impulsive/premeditated aggression are the two most-cited bimodal classifications of aggressive traits [4]. The subtypes are not mutually exclusive, and individuals may show hybrids of subtypes [30]. Online harassment research has mostly targeted harassers with instrumental aggression (i.e., premeditated, proactive aggression) who try to achieve a certain purpose (e.g., disrupting conversations, gaining infamy, or entertainment). Our work aims to take a much broader range of possible aggressive motives and traits (e.g., susceptibility to triggers) into account in characterizing harassers. To this end, we investigate whether the psychosocial factors we define above might helpfully predict self-reported harassing behavior.

3 METHOD

This paper reports the pretest results of our online survey instrument using Qualtrics. For the pretest, we used a convenient sampling of 307 participants—106 university students attending the same undergraduate course and 201 Prolific (www.prolific.co) paid subjects, mostly women, White, Democrats, and in their 20s. We plan to administer a finalized survey questionnaire to a larger, nationally representative sample. Future work will also evaluate the effect of including more psychosocial measures and develop richer, more robust models by, for example, including interactive and mediating effects.

3.1 Dependent Variable: Single-Item Screener Response

We asked participants a single-item screener question that we intended to facilitate self-reports of engaging in harassing behavior

online. We prompted: "*Sometimes, we see people post things online that we find objectionable or unacceptable. Have you ever engaged in any conflict with someone on the Internet 1) to show your disapproval?; 2) to upset them?; 3) in anger?*" We pretested these three different phrasings, all of which imply aggression and harassing behavior. We randomly assigned participants to one of the screeners using the Qualtrics randomizer: to show disapproval ($n = 102$); to upset ($n = 106$); in anger ($n = 99$). Participants responded to this screener with either "yes" or "no"—and if "no," we asked if they have ever considered engaging in online conflict ("no, but considered") or not ("no, never considered"). In this study, we set this single-item screener response as our main dependent variable for simplicity in our model.

We cross-checked the screener response with the number of harassing behaviors self-reported on a multi-item checklist. We also asked: "*There may be times when conflicts with another person on the Internet escalate. Thinking of your past conflicts with someone on the Internet, have you ever…*" and listed 16 different types of behaviors that are defined as harassing based on Lenhart et al. [18], Thomas et al. [31], and Vogels [33]. The list included: calling someone offensive names; sexual harassment; location tracking; false reporting; exposing someone to unwanted explicit content; message bombing; and incitement of dogpiling. A Mann-Whitney U test showed that there is a significant difference ($W = 7,566$, $p < .001$, $\varepsilon^2 = .07$) between people who responded "yes" versus "no" to our screener. The median number of self-reported harassing behaviors was 0 (IQR = $0 - 1$) for those who responded "no" compared to 1 (IQR = $0-2$) for those who responded "yes." A Kruskal-Wallis test indicated no significant differences between our 1) to show disapproval, 2) to upset, and 3) in anger screeners ($\chi^2(2) = 2.81$, $p = .25$).

3.2 Independent Variables: Psychosocial Measures

We measured impulsivity with the *Barratt Impulsiveness Scale-Brief* (BIS-Brief) [27], an 8-item self-report measure designed to assess a person's general impulsiveness. We scored these items (e.g., "I do things without thinking") on a 4-point scale (1 = rarely/never, 2 = occasionally, 3 = often, 4 = almost always/always) and then summed them up for a total score. The published internal consistency estimates using Cronbach's alpha for this measure range from .73 to .83; for our sample, .83.

We assessed reactive-proactive aggression with the *Reactive-Proactive Aggression Questionnaire* (RPQ) [24]. The RPQ has 11 items for assessing reactive aggression (e.g., "Reacted angrily when provoked by others") and 12 items for proactive aggression (e.g., "Had fights with others to show who was on top"). We scored these items on a 3-point scale (0 = never, 1 = sometimes, and 2 = often) and summed them to form a composite for each type. The internal consistency estimates (α) of this scale have been reported to exceed .80; for our sample, .82 (reactive aggression) and .78 (proactive aggression).

We measured impulsive-premeditated aggression with the *Aggressive Acts Questionnaire* (AAQ) [5]. The instrument typically consists of 22 items asking respondents to evaluate their four most aggressive episodes in the preceding six-month period on a 5-point ordered response scale. We employed the 5-item measure of impulsive aggression (e.g., "I lacked self-control") and the 3-item measure of premeditated aggression (e.g., "Act was planned") and asked participants to assess the most aggressive online conflict in which they had engaged. We summed scores for impulsive aggression (prior work $\alpha = .75$; for our sample, $\alpha = .98$) and premeditated aggression (prior work $\alpha = .48$; for our sample, $\alpha = .96$) respectively.

We calculated our measure of disinhibition, or beliefs about how inhibited or uninhibited people feel while interacting and engaging in certain behaviors, in the online context using the *Online Disinhibition Scale* (ODS) [32]. We used the 7-item benign disinhibition subscale (e.g., "I feel like a different person online"), excluding the 4-item toxic disinhibition subscale (e.g., "Writing insulting things online is not bullying"), because the phrasing is not negative or stigmatizing, but its score is a significant predictor of online deviant behavior. We scored the items on a 5-point ordered response scale and then summed them for a total subscale score. The published internal consistency estimates (α) of the benign disinhibition subscale is above .81; for our sample, .55.

We measured moral disengagement, limited to the online context, using the *Cyberbullying-Specific Moral Disengagement Questionnaire* (CBMDQ) [9]. The 15-item questionnaire (e.g., "Cyberbullying doesn't really hurt anyone") measures cognitive tendencies relating to minimization of harmful effects, moral justification, denial of responsibility, and dehumanization of targets. We replaced the term 'cyberbullying' with 'online harassment' for our study. We scored the items on a 5-point ordered response scale and averaged them for a composite. The internal consistency estimate (α) is reported to be .91; for our sample, .85.

We also asked participants about their gender, race, education level, and political affiliations. It is well established that online victimization occurs disproportionately across demographics [33]—except for socioeconomic status (about which there is insufficient research) and political affiliations (where both the left and right claim victimization due to political views). Perpetration of online harassment might also occur disproportionately across demographics, yet there has been little research.

3.3 Procedure

We recruited participants for this study from an undergraduate course and on the online subject pool, Prolific. Participants were shown a content warning about potentially sensitive material before the survey. Those who gave their consent to participate in the study completed an online survey questionnaire consisting of eight groups of questions. The order of the question presentation to participants was: 1) screener; 2) impulsivity (BIS-Brief); 3) reactive/proactive aggression (RPQ); 4) online disinhibition (ODS); 5) harassing behavior checklist; 6) impulsive/premeditated aggression (AAQ); 7) moral disengagement (CBMDQ); and 8) demographics. Participants took about 12 minutes to complete the survey. Upon completion, we thanked participants and compensated them for their time: for students, three course credits; for Prolific subjects, 1.75 USD.

3.4 Data Analysis

We employ R statistical packages to conduct our analysis, using a p-value of .05 for all statistical tests. To characterize online harassers, we use logistic regression to assess how the psychosocial and demographic measures predict the response to the single-item screener: *Have you ever engaged in online conflict with someone to show your disapproval; to upset them; in anger?* We fit the first set of binomial logistic regressions with the "yes" versus "no" screener response answer as the outcome variable; the second set, with "no, never considered" versus "no, but considered"; and the third set, with "no, but considered" versus "yes." After building the full models with all the measures, we rely on Akaike's Information Criteria (AIC) to select the most influential predictors.

4 RESULTS

Our sample (N = 307) consisted of 106 university students attending the same undergraduate course and 201 Prolific paid subjects, who are current U.S. residents ages 18 or older. Table 1 describes the sample. For further analyses, we limit our sample to 301 participants, excluding six participants with missing data points due to undisclosed race.

4.1 Those Who Said "Yes" vs. "No" to Having Engaged in Online Harassment

We use binomial logistic regression to analyze how psychosocial factors predict whether a person would say "yes" or "no" to our screener for having engaged in harassment in the past. First, a full logit model with all psychosocial and demographic factors included as independent variables indicates a good fit (McFadden's pseudo R^2 = .20) with 71.1% classification prediction accuracy. Table 2 summarizes the estimates of this model.

We find impulsivity, impulsive aggression, premeditated aggression, and age to be significant predictors of positive screener response. Holding other variables constant, the odds of self-reporting perpetration increases by 9% (95% CI [.01 – .17]) for each additional score in impulsivity; 8% (95% CI [.02 – .15]) for each additional score in impulsive aggression; 16% (95% CI [.01 – .33]) for each additional score in premeditated aggression; and 4% (95% CI [.00 – .07]) for each additional year in age. Inferring from the odds ratios, premeditated aggression is the strongest predictor followed by impulsivity, impulsive aggression, and age. Demographic factors except for age do not predict positive response—gender, Wald $\chi^2(2) = 3.9$, $p = .14$; race, Wald $\chi^2(3) = 1.2$, $p = .76$; political affiliation, Wald $\chi^2(4) = 3.8$, $p = .43$.

4.2 Those Who Never Imagined Harassing vs. Those Who Were Tempted

We asked those who answered "no" to the screener whether they had ever *considered* engaging in harassment. We use a separate binomial logistic regression to distinguish between those who replied "never" to those who answered "yes." We present the parsimonious logit model with influential predictors using AIC (AIC = 211.03; McFadden's pseudo R^2 = .17, prediction accuracy = 75.4%; compared to full model with McFadden's pseudo R^2 = .22, prediction accuracy = 73.3%). Table 3 summarizes this model's findings. Reactive aggression is the strongest predictor of having considered online harassment. Holding other variables constant, the odds of considering such behavior increase by 12% (95% CI [.01 – .24]) for each additional score in impulsivity; 18% (95% CI [.05 – .35]) for each additional score in reactive aggression; and 13% (95% CI [.07 – .21]) for each additional score in impulsive aggression. Meanwhile, the odds decrease by 21% (95% CI [.00 – .38]) for each additional score in proactive aggression.

4.3 Those Who Were Tempted vs. Those Who Actually Did It

We also attempt to distinguish between those who thought about but did not carry out harassing behavior and those who did engage in harassment. Again, we present a parsimonious logit model using AIC (AIC = 302.46; McFadden's pseudo R^2 = .11, prediction accuracy = 63.8%; compared to full model with McFadden's pseudo R^2 = .15, prediction accuracy = 66.4%). Table 4 summarizes the results of this model. Premeditated aggression is the strongest predictor of following through on thoughts about online harassment. Holding other variables constant, the odds of executing perpetration increase by 22% (95% CI [.07 – .40]) for each additional score in premeditated aggression and 3% (95% CI [.00 – .06]) for each additional year in age.

5 DISCUSSION AND FUTURE WORK

Our study of people who do and do not engage in behaviors commonly labeled as "online harassment" provides insight into one important facet of online conflict—what characteristics distinguish participants from nonparticipants. In this early research, we discover that psychological factors are more predictive than demographic factors in identifying people who report engaging in online harassment. Overall, impulsivity, impulsive aggression, premeditated aggression, and age predict those who self-reported engaging in conflict and those who did not. Specifically, people with higher impulsivity, reactive aggression, and impulsive aggression appear more likely to have aggressive thoughts, while higher premeditated aggression and being older predict engaging in actual harassing behavior.

The possibility that demographic factors may not be significant predictors of harassing behavior is in tension with the results of previous research that targets specific segments of the population such as active Twitter users [14, 25] or Reddit users whose content or account has been censored [13]. In these prior studies, demographic factors such as gender and political affiliation have a significant relationship with engaging in harassment. Our general population survey may dampen demographic effects specific to certain platforms. Alternatively, while our sample size is statistically sufficient, a more representative sample might paint a different picture. An important caveat, of course, is that survey work studies participants' impressions of their behavior, which might be systematically different from their actions in the real world.

One step platforms could take to address online harassment is to create more pathways for engaging interpersonal interactions online that allow people to reflect on their own states and traits. While we cannot recommend predicting personality because the

Variable		Mean (SD)	Median (IQR)	Possible Range	Total (N = 307) n	%
Impulsivity		16.56(4.11)	16(14 – 19)	8 – 32		
Reactive Aggression		6.76(3.70)	7(4 – 9)	0 – 22		
Proactive Aggression		1.51(2.25)	1(0 – 2)	0 – 24		
Impulsive Aggression		9.41(6.42)	10(5 – 14)	5 – 25		
Premeditated Aggression		4.22(2.79)	5(3 – 6)	3 – 15		
Online Disinhibition		23.31(3.77)	23(21 – 26)	7 – 35		
Moral Disengagement		1.95(.53)	1.85(1.54 – 2.32)	1 – 5		
Age		25.43(10.57)	21(19 – 27)			
Gender	Woman				206	67.1%
	Man				85	27.7%
	Other				16	5.2%
Race	White				201	65.5%
	Asian				53	17.3%
	Black				14	4.6%
	Other*				33	10.8%
	Undisclosed				6	2.0%
Education Level	High school				80	26.1%
	Some college				119	38.8%
	Associate				16	5.2%
	Bachelor				69	22.5%
	Graduate				23	7.5%
Political Affiliation	Strong Democrat				78	25.4%
	Weak Democrat				163	53.1%
	Neutral/Independent				33	10.7%
	Weak Republican				22	7.2%
	Strong Republican				11	3.6%

*Includes American Indian or Alaskan Native (1), Middle Eastern (2), and other or mixed races (30).

Table 1: Sample mean (standard deviation) and median (interquartile range) of psychosocial measures and descriptive statistics.

Variable	Coefficient	Std. Error	Odds Ratio	95% CI	p	
(Intercept)	−5.38	1.46	.01	.00 – .13	< .001	***
Impulsivity	.08	.04	1.09	1.01 – 1.17	.029	*
Reactive Aggression	−.00	.05	1.00	.90 – 1.11	.966	
Proactive Aggression	.09	.08	1.09	.94 – 1.28	.276	
Impulsive Aggression	.08	.03	1.08	1.02 – 1.15	.010	**
Premeditated Aggression	.15	.07	1.16	1.01 – 1.33	.033	*
Online Disinhibition	.02	.04	1.02	.94 – 1.10	.640	
Moral Disengagement	.29	.29	1.33	.76 – 2.34	.313	
Age	.04	.02	1.04	1.00 – 1.07	.015	*
Woman	−1.11	.65	.33	.08 – 1.14	.090	
Man	−.71	.70	.49	.12 – 1.87	.311	
White	.43	.50	1.54	.59 – 4.30	.389	
Asian	.14	.59	1.14	.36 – 3.79	.820	
Black	.53	.76	1.70	.38 – 7.77	.484	
Education.linear	−.33	.44	.72	.29 – 1.68	.452	
Strong Democrat	.57	.56	1.76	.61 – 5.39	.305	
Weak Democrat	.22	.51	1.24	.47 – 3.46	.670	
Weak Republican	.10	.71	1.10	.27 – 4.43	.888	
Strong Republican	−.96	.94	.38	.05 – 2.29	.309	

Signif. codes: * $p < .05$ ** $p < .01$ *** $p < .001$

Table 2: Binomial logistic regression model predicting the "yes" as opposed to "no" response to the screener. Reference groups: gender (other), race (other), political affiliation (neutral/independent).

Variable	Coefficient	Std. Error	Odds Ratio	95% CI	p	
(Intercept)	−2.69	.80	.07	.01 − .31	< .001	***
Impulsivity	.11	.05	1.12	1.01 − 1.24	.028	*
Reactive Aggression	.17	.06	1.18	1.05 − 1.35	.009	**
Proactive Aggression	−.24	.12	.79	.62 − 1.00	.046	*
Impulsive Aggression	.13	.03	1.13	1.07 − 1.21	< .001	***

Signif. codes: * $p < .05$ ** $p < .01$ *** $p < .001$

Table 3: Binomial logistic regression model predicting the "no, but considered" as opposed to "no, never considered" response to the screener.

Variable	Coefficient	Std. Error	Odds Ratio	95% CI	p	
(Intercept)	−3.68	.91	.03	.00 − .14	< .001	***
Impulsivity	.07	.04	1.07	1.00 − 1.15	.065	
Proactive Aggression	.11	.07	1.12	.98 − 1.28	.106	
Impulsive Aggression	.05	.03	1.05	.99 − 1.12	.092	
Premeditated Aggression	.20	.07	1.22	1.07 − 1.40	.004	**
Age	.03	.01	1.03	1.00 − 1.06	.026	*

Signif. codes: * $p < .05$ ** $p < .01$ *** $p < .001$

Table 4: Binomial logistic regression model predicting the "yes" as opposed to "no, but considered" response to the screener.

science is flimsy and may further harm marginalized groups, allowing users to reflect on their own psychosocial characteristics and choose interaction pathways to manage their own behavior might promote more and better deescalation. For example, design friction could be a useful feature for some users in some contexts; instead of imagining such frictions as "pre-punishment," we could conceive and promote them as "safety features" with small inconveniences for everyone [21]. Design interventions that adapt to people's experiences and dispositions might provide more effective remediation than the currently challenged one-size-fits-all, after-the-fact approach.

Platforms could also open up more opportunities for solutions by examining whether users' psychosocial characteristics are more likely to be traits or states—that is, are some users consistent across contexts while others vary depending on the context in which they are engaging? Understanding context-specific dispositions to engage in anger or aggression is important for designing contextually appropriate and principled remedies. For example, a user who is often aggressive and engaging in racist behaviors online is very different from a user who is aggressive because they are often combating racism. Thus, any designs that consider psychosocial dispositions must also consider those dispositions in tandem with rights and principles (e.g., civil rights).

REFERENCES

[1] Kathleen P. Allen. 2015. "We don't have bullying, but we have drama": Understandings of bullying and related constructs within the social milieu of a U.S. high school. *Journal of Human Behavior in the Social Environment* 25, 3 (April 2015), 159–181. https://doi.org/10.1080/10911359.2014.893857

[2] Craig A. Anderson and Brad J. Bushman. 2002. Human aggression. *Annual Review of Psychology* 53, 1 (2002), 27–51. Publisher: Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA.

[3] Donald A. Andrews, James Bonta, and J. Stephen Wormith. 2011. The risk-need-responsivity (RNR) model: Does adding the good lives model contribute to effective crime prevention? *Criminal Justice and Behavior* 38, 7 (2011), 735–755. Publisher: Sage Publications Sage CA: Los Angeles, CA.

[4] Julia C. Babcock, Andra L. T. Tharp, Carla Sharp, Whitney Heppner, and Matthew S. Stanford. 2014. Similarities and differences in impulsive/premeditated and reactive/proactive bimodal classifications of aggression. *Aggression and Violent Behavior* 19, 3 (May 2014), 251–262. https://doi.org/10.1016/j.avb.2014.04.002

[5] Ernest S. Barratt, Matthew S. Stanford, Lynn Dowdy, Michele J. Liebman, and Thomas A. Kent. 1999. Impulsive and premeditated aggression: A factor analysis of self-reported acts. *Psychiatry Research* 86, 2 (May 1999), 163–173. https://doi.org/10.1016/S0165-1781(99)00024-4

[6] Lindsay Blackwell, Mark Handel, Sarah T. Roberts, Amy Bruckman, and Kimberly Voll. 2018. Understanding "bad actors" online. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–7. https://doi.org/10.1145/3170427.3170610

[7] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, Portland Oregon USA, 1217–1230. https://doi.org/10.1145/2998181.2998213

[8] Naomi Craker and Evita March. 2016. The dark side of Facebook®: The Dark Tetrad, negative social potency, and trolling behaviours. *Personality and Individual Differences* 102 (2016), 79–84. Publisher: Elsevier.

[9] Samantha Day and Lambros Lazuras. 2016. The Cyberbullying-Specific Moral Disengagement Questionnaire (CBMDQ-15). (2016). http://shura.shu.ac.uk/12890/

[10] Dominick DeMarsico, Nadia Bounoua, Rickie Miglin, and Naomi Sadeh. 2021. Aggression in the digital era: Assessing the validity of the cyber motivations for aggression and deviance scale. *Assessment* (Feb. 2021), 107319112199008. https://doi.org/10.1177/1073191121990088

[11] Stelios N. Georgiou, Kyriakos Charalambous, and Panayotis Stavrinides. 2020. Mindfulness, impulsivity, and moral disengagement as parameters of bullying and victimization at school. *Aggressive behavior* 46, 1 (2020), 107–115. Publisher: Wiley Online Library.

[12] Eveline Gutzwiller-Helfenfinger. 2015. Moral disengagement and aggression: Comments on the special issue. *Merrill-Palmer Quarterly* 61, 1 (2015), 192–211. Publisher: JSTOR.

[13] Oliver L. Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–35. https://doi.org/10.1145/3479610

[14] Yiqing Hua, Mor Naaman, and Thomas Ristenpart. 2020. Characterizing Twitter users who engage in adversarial interactions against political candidates. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–13. https://doi.org/10.1145/3313831.3376548

[15] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)* 25, 2 (2018), 1–33.

[16] Robin M. Kowalski, Gary W. Giumetti, Amber N. Schroeder, and Micah R. Lattanner. 2014. Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. *Psychological Bulletin* 140, 4 (July 2014),

1073–1137. https://doi.org/10.1037/a0035618

[17] Anna Kurek, Paul E. Jose, and Jaimee Stuart. 2019. 'I did it for the LULZ': How the dark personality predicts online disinhibition and aggressive online behavior in adolescence. *Computers in Human Behavior* 98 (Sept. 2019), 31–40. https://doi.org/10.1016/j.chb.2019.03.027

[18] Amanda Lenhart, Michele Ybarra, Kathryn Zickuhr, and Myeshia Price-Feeney. 2016. *Online harassment, digital abuse, and cyberstalking in America.* Data and Society Research Institute.

[19] Alice Marwick and danah boyd. 2014. 'It's just drama': Teen perspectives on conflict and aggression in a networked era. *Journal of Youth Studies* 17, 9 (Oct. 2014), 1187–1204. https://doi.org/10.1080/13676261.2014.901493

[20] Alice E. Marwick. 2021. Morally motivated networked harassment as normative reinforcement. *Social Media+ Society* 7, 2 (2021), 20563051211021378.

[21] Sahar Massachi. 2021. How to save our social media by treating it like a city. https://www.technologyreview.com/2021/12/20/1042709/how-to-save-social-media-treat-it-like-a-city/

[22] Joshua D. Miller, Amos Zeichner, and Lauren F. Wilson. 2012. Personality correlates of aggression: Evidence from measures of the five-factor model, UPPS model of impulsivity, and BIS/BAS. *Journal of Interpersonal Violence* 27, 14 (2012), 2903–2919.

[23] Catherine Page Jeffery. 2021. '[Cyber]bullying is too strong a word…': Parental accounts of their children's experiences of online conflict and relational aggression. *Media International Australia* (Oct. 2021), 1329878X2110485. https://doi.org/10.1177/1329878X211048512

[24] Adrian Raine, Kenneth Dodge, Rolf Loeber, Lisa Gatzke-Kopp, Don Lynam, Chandra Reynolds, Magda Stouthamer-Loeber, and Jianghong Liu. 2006. The reactive–proactive aggression questionnaire: Differential correlates of reactive and proactive aggression in adolescent boys. *Aggressive Behavior* 32, 2 (April 2006), 159–171. https://doi.org/10.1002/ab.20115

[25] Jennifer D. Rubin, Lindsay Blackwell, and Terri D. Conley. 2020. Fragile masculinity: Men, gender, and online harassment. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[26] Yunya Song, Qinyun Lin, K. Hazel Kwon, Christine H. Y. Choy, and Ran Xu. 2022. Contagion of offensive speech online: An interactional analysis of political swearing. *Computers in Human Behavior* 127 (Feb. 2022), 107046. https://doi.org/10.1016/j.chb.2021.107046

[27] Lynne Steinberg, Carla Sharp, Matthew S. Stanford, and Andra Teten Tharp. 2013. New tricks for an old measure: The development of the Barratt Impulsiveness Scale–Brief (BIS-Brief). *Psychological Assessment* 25, 1 (March 2013), 216–226. https://doi.org/10.1037/a0030550

[28] Joachim Stöber. 2001. The Social Desirability Scale-17 (SDS-17): Convergent validity, discriminant validity, and relationship with age. *European Journal of Psychological Assessment* 17, 3 (2001), 222. Publisher: Hogrefe & Huber Publishers.

[29] Randal W. Summers. 2016. *Social psychology: How other people influence our thoughts and actions.* Vol. 1. ABC-CLIO.

[30] Andra L. Teten Tharp, Carla Sharp, Matthew S. Stanford, Sarah L. Lake, Adrian Raine, and Thomas A. Kent. 2011. Correspondence of aggressive behavior classifications among young adults using the Impulsive Premeditated Aggression Scale and the Reactive Proactive Questionnaire. *Personality and Individual Differences* 50, 2 (Jan. 2011), 279–285. https://doi.org/10.1016/j.paid.2010.10.003

[31] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, et al. 2021. Sok: Hate, harassment, and the changing landscape of online abuse. (2021).

[32] Reinis Udris. 2014. Cyberbullying among high school students in Japan: Development and validation of the Online Disinhibition Scale. *Computers in Human Behavior* 41 (Dec. 2014), 253–261. https://doi.org/10.1016/j.chb.2014.09.036

[33] Emily A. Vogels. 2021. The state of online harassment. *Pew Research Center* 13 (2021).