Internet Trolling and Everyday Sadism:

Parallel Effects on Pain Perception and Moral Judgment

Erin E. Buckels

University of British Columbia

Paul D. Trapnell and Tamara Andjelovic

University of Winnipeg

&

Delroy L. Paulhus

University of British Columbia

Draft date: February 22, 2018

Abstract

Objective: To clarify the association between online trolling and sadistic personality; and provide evidence that the reward and rationalization processes at work in sadism are likewise manifest in online trolling. **Method:** Online respondents (total $N = 1{,}715$) completed self-report measures of personality and trolling behavior. They subsequently engaged in one of two judgment tasks. In Study 1, respondents viewed stimuli depicting scenes of emotional/physical suffering, and provided ratings of (a) perceived pain intensity and (b) pleasure experienced while viewing the photos. In Study 2, the iTroll questionnaire was developed and validated. It was then administered alongside a moral judgment task. **Results:** Across both studies, online trolling was strongly associated with a sadistic personality profile. Moreover, sadism and trolling predicted identical patterns of pleasure and harm minimization. The incremental contribution of sadism was sustained even when controlling for broader antisocial tendencies (i.e., the Dark Triad, callous-emotionality, and trait aggression). **Conclusion:** Results confirm that online trolling is motivated (at least in part) by sadistic tendencies. Coupled with effective rationalization mechanisms, sadistic pleasure can be consummated in such everyday behaviors as online trolling.

Keywords: Online trolling, cyber-deviance, Dark Triad, Dark Tetrad, everyday sadism

Internet Trolling and Everyday Sadism:

Parallel Effects on Pain Perception and Moral Judgment

In March 2016, Microsoft released a chatbot named "Tay" into the Twitter-sphere (Dewey, 2016). They touted this advance in artificial intelligence as a realistic emulation of a teenage girl, with the capacity to become smarter over time through machine learning technology. Unfortunately, Tay's ability to learn from online interactions proved to be her downfall. Tay was bombarded with the worst of the Internet, courtesy of her fellow Twitter users. This intensive education prompted Tay to spout a series of racist and anti-Semitic messages that forced Microsoft to silence her after only 16 hours online (Rodriguez, 2016). The incident left many bewildered at the emergence of a "Nazi intelligence" from the cyber-collective, rather than a progressive social intellect. Why did the project fail so miserably?

We suggest that Tay embodied supreme troll-bait: An unprecedented lure for the so-called "Twitter trolls" to corrupt a scientific advance for their own amusement (see Ohlheiser, 2016). Online trolling has only become widely-known over the past decade[1] (Maltby et al., 2016) and is proliferated across all social media by a small but energetic minority (Gammon, 2014). Trolling behaviors are diverse and continue to evolve (Fichman & Sanfillippo, 2016; Phillips, 2015); they exemplify the tradeoff between prosocial and antisocial aspects of Internet culture. Depending on whom you ask, trolls attempt to [challenge / correct / educate] or [abuse / antagonize / deceive / harm] other Internet users. Due to such shifting descriptions, trolling is often confused with cyber-bullying and other forms of online misbehavior. But the latter can, in fact, be distinguished from trolling based on form, content, intent, and consequence (Fichman & Sanfillippo, 2016; Hardaker, 2010; Leone, 2017; Shachaf & Hara, 2010).

[1] The phenomenon can be traced back to 'flaming' in 1990s messaging boards

To avoid these ambiguities, we follow Fichman and Sanfillippo's (2016) definition of trolling as an intentionally disruptive behavior that occurs (a) in the context of Internet discourse, and (b) among users having no existing relationship in real life. Trolls may act alone or with others; either indiscriminately, or selectively toward certain individuals, social groups, political parties, corporate entities, and so on. In some cases, trolling reduces to a simple-minded attempt to annoy others. In other cases, trolling demonstrates considerable skill, creativity, dedication, and perseverance; and as such, the behavior has even been likened to an art form (Dynel, 2016; Leone, 2017).

We appreciate Fichman and Sanfillippo's (2016) definition for its acknowledgement that trolling tendencies vary in kind and degree. Not all Twitter users approached Tay with an intention to corrupt (yet some did). Likewise, not all Internet users seek to disrupt online discourse (yet some do). Clearly, there are individual differences at work. Indeed, recent empirical research (Buckels, Trapnell, & Paulhus, 2014; Craker & March, 2016; March, Grieve, Marrington, & Jonason, 2017; Sest & March, 2017) revealed links between trolling and sadistic personality traits (Buckels, Jones, & Paulhus, 2013) and sensitivity to antisocial rewards (Foulkes, Viding, McCrory, & Neumann, 2014). That link with sadism also holds for other past-times involving virtual cruelty: For example, violent video gaming (Greitemeyer, 2015; Greitemeyer, Sagioglou, 2017) and cyberbullying (van Geel, Goemans, Toprak, & Vedder, 2017). Together, these findings point to an appetitive motivation in which the instigating or observing of others' distress is rewarding (cf. Foulkes, McCrory, Neumann, & Viding, 2014; Schumpe & Lafrenière, 2016), thereby perpetuating trolling behaviors. For hardcore trolls, the hunt for *lulz* [2] may simply be too tantalizing to resist.

We have argued elsewhere that, along with the appetitive enjoyment of cruelty, sadistic

[2] Lulz is an Internet neologism meaning 'thrills' or 'kicks'.

behavior requires callousness, that is, a deficit in empathy (Buckels et al., 2013; Paulhus, 2014). Any psychological mechanism that inhibits empathy would help release the sadistic impulses of those who are rewarded by cruelty (Nell, 2006). The anonymity of the Internet minimizes a powerful deterrent, that is, social repercussions (Suler, 2004). Hence, it is the ideal venue to unleash latent malevolence. Much like the widespread enjoyment of violent media (e.g., sports, film, and video games), recreational trolling is now common and, for the most part, socially accepted. For that reason, we apply the term *everyday sadism* [3] (Paulhus & Dutton, 2014).

There is undoubtedly some pleasure to be gained from a clever riposte or jab at an irritating target online. For individuals with sadistic personalities, that pleasure is sufficiently motivating to engender harm to undeserving targets (Buckels et al., 2013; Pfattheicher, Keller, & Knezevic, 2017). Cruel behavior is enabled by disinhibiting empathy and downplaying repercussions. With increasing access to the Internet, everyday sadists can easily indulge their appetite for cruelty. This combination of (heightened) appetitive and (diminished) avoidance processes may help explain why online trolls behave the way they do.

Overview

We conducted two studies to evaluate the association of sadistic personality with online trolling and reveal the psychological processes of reward and rationalization that are operating in both sadism and online trolling. In Study 1, we evaluated responses to images of physical and emotional pain. In Study 2, we evaluated a possible moral deficiency using an unobtrusive index of harm rationalization. We expected to find evidence that sadism and trolling exhibit parallel patterns of pleasure responses and biased judgments of harm/suffering. Data are publicly available on the Open Science Framework (https://osf.io/gtnkj).

[3] This term, *everyday sadism*, has proved useful in distinguishing criminal or sexual sadism from more common and socially acceptable forms.

Study 1: Online Trolling and Sadistic Pain Perception

Study 1 examined how trolls and sadists evaluate pain intensity in visual depictions of others' suffering (i.e., photos of injuries, physical harm, and facial expressions of pain). Although suspecting a systematic bias in trolls' and sadists' evaluations of these stimuli, we had to consider competing hypotheses about the direction of that bias. First, given their frequent need to justify harm, trolls and sadists may show a habitual tendency to downplay the extent of others' suffering. This hypothesis suggests a *negative* association between trolling/sadism scores and perceived pain intensity (**H1a**). The alternative hypothesis is that trolls and sadists tend to exaggerate others' suffering (i.e., engage in sadistic fantasy) to maximize their pleasure. This hypothesis suggests a *positive* association between trolling/sadism scores and perceived pain intensity (**H1b**).

Study 1 also sought to identify individuals who derive pleasure from others' suffering: We predicted that both sadism and trolling would be positively associated with pleasure ratings (**H2**). Finally, following Buckels et al. (2014), we expected pleasure ratings to act as a mediator of any trolling/sadism effects on pain perception (**H3**). To rule out a variety of alternatives, we included measures of the so-called *Dark Triad* of personality: That is, subclinical psychopathy, narcissism, and Machiavellianism (Paulhus & Williams, 2002)[4]. We tested our hypotheses in an online experiment with a large sample of community adults.

Method

Participants. Participants were 345 adults (48.2% male) recruited on Amazon's Mechanical Turk website to complete an online survey in exchange for $0.50. Mean age was 34.4 years (*SD* = 12.69). Participant location was limited to the United States and the most

[4] When considered in tandem with sadism, this suite of malevolent traits is known as the *Dark Tetrad* of personality (Chabrol, Van Leeuwen, Rodgers, & Séjourné, 2009; Paulhus, 2014).

frequent ethnic ancestries reported in this sample were Caucasian (78.63%), African American (7.2%), Asian (6.9%), Hispanic or Latino (3.9%), and various other ethnicities (1.6%); 1.6% declined to answer. When asked about their highest level of educational attainment, 14.8% indicated high school, 41.1% had some college or university, 36.2% held an undergraduate degree, 7.2% had graduate education, and 1% declined to answer.

Measures. The online measures were divided into three sections: (1) personality questionnaires, (2) scales for rating the images of suffering, and (3) self-reports of trolling tendencies.

Personality Questionnaires. The 27-item Short Dark Triad scale (Jones & Paulhus, 2014) was used to assess narcissism (e.g., "*I have been compared to famous people*," α = .76), Machiavellianism (e.g., "*It's not wise to tell your secrets*," α = .82), and subclinical psychopathy (e.g., "*People often say I'm out of control*," α = .82), with nine items per subscale. We used the 18-item Comprehensive Assessment of Sadistic Tendencies (CAST; Buckels & Paulhus, 2013; α = .89 in this sample) to assess sadistic personality. Example items include, "*I enjoy physically hurting people*" and "*I enjoy making jokes at the expense of others*," as rated on a 5-point scale from 1 (*Strongly disagree*) to 5 (*Strongly agree*).

Pain perception and pleasure ratings. As part of the online questionnaire, participants were presented with six photographs depicting people in various degrees of physical or emotional pain. These photographs were selected from the International Affective Picture System (Lang, Bradley, & Cuthbert, 2005; picture codes: 2399, 2900, 3053, 3150, 8480, 9402). Participants viewed the photos at a pace of their choosing. Two ratings were requested for each photo as it was displayed on the screen: (1) perceived pain intensity ("*How much pain is this person in?*"), from 1 (*no pain*) to 5 (*severe pain*), and (2) obtained pleasure ("*How pleasing [or*

unpleasing] is this photo?"), from 1 (*very unpleasing*) to 7 (*very pleasing*). Pain intensity (M = 3.70, SD = 0.44) and pleasure (M = 2.19, SD = 0.73) ratings were standardized and composite scores were computed as the mean of standardized ratings across the stimuli (α's = .39 for pain intensity[5] and .68 for pleasure, respectively).

Trolling tendencies. Four items (α = .85) assessed trolling identity and behavior (Buckels et al., 2014): "*I have directed people to shock websites for the lulz*," "*I like to troll people in forums or the comments section of websites*," "*The more beautiful and pure a thing is, the more satisfying it is to corrupt*," and "*I enjoy griefing other players in multiplayer games*." Items were rated on a 5-point scale from 1 (*Strongly disagree*) to 5 (*Strongly agree*).

Results

Characteristics of online trolls. Overall endorsement of the trolling items was relatively low (M = 1.61, SD = .86), but scores spanned the entire range of the scale (min = 1.00; max = 5.00). Younger participants had higher trolling scores than did older participants (r = -.24, p < .001); and men (M = 1.84, SD = 0.95) scored higher than did women (M = 1.39, SD = 0.67), t(255.41) = 4.76, p < .001, d = .60.

As seen in Table 1, online trolling was positively associated (r's > .25) with scores on all Dark Tetrad measures. Results from a multiple regression analysis (also displayed in Table 1) indicated that sadism maintained a strong positive association with trolling even after controlling for scores on the Dark Triad measures. Psychopathy also maintained a significant (though weaker) unique association with trolling. The pattern of association generally replicates that found by Buckels and colleagues (2014).

[5] This low alpha undoubtedly reflects the diversity of the stimuli. For simplicity, correlations with the composite pain intensity scores are reported in the results that follow, but the sadism and trolling effects were also significant using Hierarchical Linear Modeling (with subject entered as a random effects variable and photo set as a random effects variable with six levels).

Pain perception and obtained pleasure. Table 2 presents bivariate correlations with the composite pain perception and pleasure ratings. Patterns of association were highly similar across measures of online trolling, sadism, psychopathy, and Machiavellianism: Specifically, scores on these four predictor variables were (a) negatively associated with perceived pain intensity and (b) positively associated with self-reported pleasure from others' pain.

In addition, pain and pleasure ratings were separately regressed on the Dark Tetrad. Those analyses revealed that sadism was a unique predictor of pain (β = -.25, p = .006) and pleasure (β = .35, p < .001) ratings. Psychopathy was also a unique (though lesser) predictor of pleasure (β = .17, p = .049), but not pain ratings (β = -.03, p = .79). No other associations were significant (p's > .59).

Follow-up analyses. Follow-up mediation analyses were conducted using PROCESS for SPSS (v. 2.15; Hayes, 2016). All variables were standardized prior to entry. Significance was tested both with Sobel tests, and with bootstrapped 95% confidence intervals for the standardized indirect effects (*ab*) constructed with 10,000 resamples and a percentile distribution. The first mediation model (see Figure 1; Model A) examined pleasure as a possible mediator of sadism's negative association with perceived pain intensity. The mediated effect of sadism via pleasure was significant, 95% CI for *ab* = [-.27, -.11], Sobel's z = -5.30, p < .001; whereas the direct effect (*c'*) was not significant, $t(301)$ = -1.64, p = .10, indicating full mediation[6]. An identical pattern emerged for trolling (see Figure 1; Model B); the mediated effect of trolling via pleasure was significant, 95% CI for *ab* = [-.24, -.09], Sobel's z = -5.05, p < .001; and the direct effect (*c'*)

[6] The mediation pattern did not change when psychopathy was entered as a covariate. Sadism was still positively associated with pleasure from others' pain when controlling for psychopathy (β = .35, p < .001). In turn, pleasure remained negatively associated with perceived pain intensity when controlling for both sadism and psychopathy (β = .31, p < .001). The indirect effect was significant, 95% CI for *ab* = [-.23, -.06], Sobel's z = -3.57, p < .001. In contrast, the direct effect of sadism (*c'*) was not significant (β = .12, p = .18), nor was the effect of psychopathy when controlling for sadism and pleasure (β = .02, p = .78).

was not significant, $t(301) = -1.84$, $p = .07$.

We conducted alternative mediation analyses with perceived pain intensity as a mediator of the relationship between sadism/trolling and pleasure. These analyses indicated that, although significant, the indirect effects via perceived pain intensity were weaker in magnitude and only met the criteria for partial mediation: Perceived pain intensity was negatively associated with pleasure when controlling for sadism scores, $\beta = -.33$, $p < .001$, and the indirect effect via perceived pain intensity was significant, 95% CI for $ab = [.04, .15]$, Sobel's $z = 3.92$, $p < .001$, but so too was the direct effect (c') of sadism, $\beta = .38$, $p < .001$. Similarly, perceived pain intensity was negatively associated with pleasure when controlling for trolling scores, $\beta = -.35$, $p < .001$, and the indirect effect via perceived pain intensity was significant, 95% CI for $ab = [.04, .15]$, Sobel's $z = 3.81$, $p < .001$, but so too was the direct effect (c') of trolling ($\beta = .31$, $p < .001$), indicating partial mediation.

Discussion

Results from Study 1 replicate and extend our previous research (Buckels et al., 2014). In particular, we found additional evidence that online trolling is an instance of everyday sadism: Trolling tendencies were accompanied by a sadistic personality profile. In further support of that association, we found parallel effects of sadism and trolling in participants' reactions to photos depicting suffering. Both trolls and sadists were pleased by visual representations of people in physical/emotional pain. At the same time, they appeared to downplay the magnitude of that pain (a pattern that supports hypothesis H1a over H1b). These findings are consistent with emerging evidence that sadists engage in psychological rationalization processes to avoid guilt (Trémolière & Djeriouat, 2016).

Finally, sadism and trolling yielded parallel mediation patterns: In both cases, pleasure

reactions significantly mediated the associations with perceived pain intensity. The pattern suggests that trolls and sadists underestimate others' pain *because* they find it pleasurable. In short, trolling tendencies show evidence of an appetite for cruelty and a tendency to underestimate others' suffering. These findings offer a glimpse into the mind of online trolls and further confirm the sadistic motivations behind their behavior.

Study 2: Online Trolling, Sadism, and Morality

Previous empirical research suggests that the tendency to underestimate others' suffering reflects an automatic perceptual process (Craig, Versloot, Goubert, Vervoort, & Crombez, 2010), one that is influenced by previous experience with inflicting pain on others (e.g., in medical contexts; Cheng, Lin, Liu, Hsu, Lim, Hung, & Decety, 2007). We wondered whether rationalization of harming others might also operate at a higher, more controlled level. That level would better implicate moral culpability. To this end, we drew on emerging evidence that antisocial traits such as sadism (Trémolière & Djeriouat, 2016) and psychopathy (Marshall, Watts, & Lilienfeld, 2018) are accompanied by deficient moral judgment. Psychological rationalization may help explain how otherwise average people can reconcile cruel behavior with a positive self-view.

The study by Trémolière and Djeriouat (2016) showed that, unlike non-sadists (who typically use negative emotions such as disgust and anger to guide their moral judgments), those high in sadism use *positive emotions* (e.g., amusement, excitement) in judging right from wrong. By this standard, behavioral transgressions that are pleasing/funny are judged morally acceptable and deserving of leniency, while unexciting/boring transgressions are unacceptable and punishable. In other words, sadistic moral judgments are guided by pleasure, not pain.

Accordingly, in Study 2, we attempted to replicate and extend previous findings

involving sadism and moral judgment. Given the evidence that trolling is associated with sadistic appetite plus rationalization (confirmed in Study 1), we hypothesized that trolls would display a deviant pattern of moral reasoning: A pattern resembling the pleasure-driven moral compass of individuals scoring high on a trait measure of sadism. Specifically, we expected sadism and trolling to predict lenient moral judgments (**H1**) and greater use of positive affect in moral decision-making (**H2**). Finally, as in Study 1, we expected parallel mediation patterns for sadism and trolling: Positive affect should emerge as a statistical mediator of any sadism/trolling effects on moral judgment (**H3**). We tested these predictions in an online experiment with a large undergraduate student sample.

To better enable this study, we developed and validated a more comprehensive measure of trolling identification and behavior (using a separate, large sample of university students). The final 12-item scale was labeled the *iTroll* questionnaire. As in Study 1, we also assessed a broad range of dark personalities to confirm the specific profile of the trolling personality. Here, we added trait aggression (Buss & Perry, 1992) and callous/unemotional traits (Frick & Ray, 2015) to the personality package.

Method

Participants and procedure. Participants were introductory psychology students who completed an online survey in exchange for partial course credit. Sample A consisted of 1134 respondents (800 women, 326 men; 8 did not specify). Median age was 18 years old, with 93.1% of the participants under the age of 25. The most frequent ethnic ancestries reported in this sample were European (58.1%), other (12.8%, most being combined Euro and non-Euro identifications), East Asian (7.9%), aboriginal (6.7%), South Asian (4.2%), African (2.7%), and Pacific Islands (2.2%).

Sample B consisted of 236 respondents (174 women, 62 men). Median age was 19 years old, with 95% of the participants under the age of 25. The most frequent ethnic ancestries reported in this sample were European (59%), other (12.8%, most being combined Euro and non-Euro identifications), South Asian (6.7%), Indigenous (4.9%) and African (4.9%), East Asian (3.7%), and Pacific Islands (3.7%).

Measures (Sample A). Sadistic personality was assessed with the 18-item Comprehensive Assessment of Sadistic Tendencies (CAST; Buckels & Paulhus, 2013; α = .89 in this sample). To construct a more comprehensive and reliable measure of online trolling tendencies, we generated a pool of 18 self-descriptive statements capturing enjoyment and participation in trolling behavior (e.g. "*I enjoy trolling other people*"), attitudes toward trolling (e.g., "*Trolling behavior should be punished*" [R]), and identification with trolling culture (e.g., "*I identify with trolling culture*"). Participants responded to these statements using a 5-point rating scale from 1 (*Strongly disagree*) to 5 (*Strongly agree*). The resulting data were used to identify and discard six deficient items (on the basis of low endorsement rates, an initial principal components analysis, and content considerations). The 12 selected items that compose the final iTroll questionnaire (α = .91 in this sample) are presented in the online supplemental materials.

Trolling validity measures. We asked participants to rate their level of enjoyment when trolling different targets (or imagine what it might be, if they had never trolled that type of target). Targets included, "*General public / strangers*," "*Other trolls*," "*Corporations*," "*Celebrities*," and "*People who you know in real life (offline).*" A trolling enjoyment variable was computed as the mean of the standardized scores for the five targets (α = .87). Participants also estimated hours per week spent (a) on the Internet (all activities), (b) using social media (e.g., Facebook, Twitter), and (c) trolling; the latter rating was used as an index of trolling

frequency. Finally, self-perceived competence at trolling was assessed with the item, "*How skilled are you at trolling?*" on a 5-point scale from 1 (*novice*) to 5 (*advanced/master*).

Measures (Sample B).

Individual difference measures. Sadistic personality was assessed with the 10-item Short Sadistic Impulse Scale (SSIS; O'Meara et al., 2011; α = .88), which correlates highly with the CAST (typical *r*'s > .70); an example item is, *"Hurting people would be exciting,"* as rated on a 5-point scale from 1 (*Strongly disagree*) to 5 (*Strongly agree*). Subclinical psychopathy was indexed with nine items from the Short Dark Triad questionnaire (SD3; Jones & Paulhus, 2014; α = .77), rated on 5-point scales from 1 (*Strongly disagree*) to 5 (*Strongly agree*).

Trait aggression was indexed via a short-form version of Buss and Perry's (1992) Brief Aggression Questionnaire, developed by Webster et al. (2014; 18 items, e.g., *"If someone hits me, I hit back;"* α = .86); items were rated on 5-point scales from 1 (*Extremely uncharacteristic of me*) to 5 (*Extremely characteristic of me*). Callous-unemotional traits were assessed via a short-form of the Inventory of Callous-Unemotional Traits (Ray, Frick, Thornton, Steinberg, & Cauffman, 2016; 10 items, e.g., *"I feel bad/guilty when I do something wrong;"* α = .83) using a 4-point rating scale from 0 (*not at all true*) to 3 (*definitely true*). Online trolling tendencies were assessed by the iTroll questionnaire (12 items, e.g., "*I enjoy trolling other people*;" α = .86), which was developed and validated with Sample A (as detailed above).

Moral judgment and affect. Participants responded to Trémolière and Djeriouat's (2016) moral judgment items, which were based on prior work by Cushman (2008). Their measure comprises nine social scenarios that are relevant to three central determinants of moral judgment (intentionality, causality, and harm) in three different contexts (intentional harm, attempted harm, and accidental harm); see the online supplemental materials. Participants were randomly

assigned to view three of the nine possible scenarios (including one for each harm variety). Participants provided three ratings for each scenario: (1) wrongness ("*Was the perpetrator morally wrong in this situation?*"), (2) guilt ("*Should the perpetrator feel guilty in this situation?*"), and (3) punishment ("*Does the perpetrator deserve punishment in this situation?*"), as rated on 5-point scales from 1 (*No*) to 5 (*Yes*). Because the wrongness, guilt, and punishment ratings were positively correlated ($.39 < r\text{'s} < .80$), they were standardized and combined into a composite score to index *perceived perpetrator culpability* (Cronbach's α's = .87, .77, and .71 for the intentional, attempted, and accidental harm conditions, respectively). Higher scores represent harsher judgments.

Affect produced by the scenarios (and use of affect in the moral decision-making process) was assessed via a 10-item affect rating scale used by Trémolière and Djeriouat (2016). The key items were interspersed with filler items from the Positive and Negative Affect Schedule (Watson, Clark & Tellegen, 1988). Participants rated the extent to which these feelings affected their judgment for each scenario. Five items assessed positive affect: *enthusiastic, delighted, excited, cheerful,* and *joyful* (α's > .87 across the intentional, attempted, and accidental scenarios). An additional five items assessed negative affect: *sad, disgusted, outraged, downhearted,* and *loathing* (α's > .67 across the intentional, attempted, and accidental scenarios).

Results

iTroll psychometric evaluation and validation. Internal consistency estimates for iTroll total scores were good in both samples (α's > .86). In Sample A, alphas were high for men and women (both α's = .90). Mean iTroll scores were 2.27 (SD = 0.73), with scores being higher among men (M = 2.55, SD = 0.80) than women (M = 2.14, SD = 0.67), $t(483.19) = 8.01$, $p < .001$, $d = .73$.

To examine the factor structure of the iTroll scale, we subjected item scores in Sample A to principal axis factoring with an oblique (direct oblimin) rotation. This analysis identified two factors with eigenvalues greater than one (6.0 and 2.17, accounting for 50.0% and 18.1% of the variance, respectively). As the second iTroll factor was composed entirely of con-trait (reversed scored) items, the two-factor structure was interpreted as an artifact of item wording. Indeed, when a single factor was requested, all items loaded > .47 on that factor.

Bivariate correlations between iTroll scores, sadism, and the validity measures in Sample A are displayed in Table 3. Across the entire sample, iTroll scores were positively correlated with enjoyment of trolling, trolling frequency, and perceived trolling skill—consistent with the view that these measures tap a common trolling construct. Sadism scores were also positively correlated with all trolling measures (i.e., iTroll, trolling enjoyment, trolling frequency, and perceived trolling skill).

As displayed in Table 1, iTroll scores in Sample B were significantly positively associated with those of sadism, aggression, callous-unemotionality, and psychopathy, *r*'s > .30. Results from a multiple regression analysis (also displayed in Table 1) with sadism, aggression, callous-unemotionality, and psychopathy predicting iTroll scores revealed that sadism and psychopathy were each unique predictors of online trolling. In sum, the iTroll questionnaire is a psychometrically-sound measure of trolling tendencies that displays the expected associations with trolling variables and dark personality traits. Having established sufficient construct validity for the iTroll measure, we next turned to the culpability judgments.

Moral culpability. Moral culpability scores were strongly and positively correlated across the intentional harm and attempted harm conditions ($r = .64$, $p < .001$), indicating that participants who judged perpetrators of intentional harm more harshly were similarly harsh

toward perpetrators of attempted harm. In contrast, culpability scores for the attempted and accidental harm conditions were, overall, uncorrelated (r = -.13, p = .06). Interestingly, there was a significant negative correlation between culpability scores for intentional and accidental harm (r = -.19, p = .005), indicating that participants who judged accidental harm more harshly were slightly more lenient of intentional harm compared to others, and vice versa. Most important, trolling and sadism scores were negatively correlated with culpability judgments in both the intentional and attempted harm conditions (see Table 4).

Negative affect. As expected, use of negative affect differed across the three harm conditions, $F(2, 410) = 195.45$, $p < .001$. Pairwise comparisons (with Bonferroni adjustments) indicated that use of negative affect was greatest for intentional harm (M = 15.52, SD = 4.62), followed by attempted harm (M = 13.99, SD = 4.62) and accidental harm (M = 9.99, SD = 3.68), with all comparisons significant, p's < .001. Correlational analyses confirmed the important role of negative affect in moral decision-making: Use of negative affect was positively correlated with culpability scores in the intentional (r = .25, $p < .001$), attempted (r = .29, $p < .001$), and accidental (r = .30, $p < .001$) harm conditions. Overall, participants made harsher judgments when negative affect levels were high, and more lenient judgments when negative affect levels were low.

Positive affect. As was the case for negative affect, use of positive affect differed across the three harm conditions, $F(2, 442) = 7.86$, $p < .001$. Pairwise comparisons (with Bonferroni adjustments) indicated that positive affect was strongest in the attempted harm condition (M = 7.42, SD = 4.29), as compared to the intentional (M = 6.76, SD = 3.44) and accidental harm (M = 6.70, SD = 3.33) conditions, p's = .007 and .005, respectively. The latter two conditions did not differ on positive affect, $p > .99$. Correlational analyses indicated that use of positive affect was

significantly negatively associated with culpability scores in the intentional (r = -.46, p < .001) and attempted (r = -.30, p < .001) harm conditions (but not in the accidental harm condition, r = .09, p = .18). In other words, participants who reported greater use of positive affect in their moral decision-making were more lenient toward intentional and attempted harm than were others.

Individual differences also influenced the use of positive affect in culpability judgments. These critical results are summarized in Table 5. Across all conditions (intentional, attempted, and accidental harm), higher trolling, sadism, and psychopathy scores were correlated with stronger positive affect ratings. Results from a series of multiple regressions (also displayed in Table 5), indicated that sadism and trolling were significant unique predictors of positive affect in moral decision-making; in contrast, psychopathy, aggression, and callous-unemotionality failed to reach significance when controlling for overlap with the other measures.

Follow-up analyses (intentional harm). As a final test of the sadism hypothesis, we examined mediation by positive affect in the intentional harm condition. Significance was tested both with Sobel's tests and bootstrapped 95% confidence intervals for the standardized indirect effects (*ab*), which were constructed with 10,000 resamples and a percentile distribution. The first mediation model (see Figure 2, Model A) examined positive affect as a mediator of sadism's negative association with culpability scores. The mediated effect of sadism via positive affect was significant, 95% CI for *ab* = [-.25, -.08], Sobel's z = -4.38, p < .001. The direct effect of sadism on culpability judgments (*c'*) was not significant, $t(221)$ = -1.94, p = .054, which indicates full mediation.

The second mediational model (see Figure 2, Model B) examined positive affect as a mediator of trolling's negative association with culpability scores. The mediated effect of trolling

via positive affect was significant, 95% CI for *ab* = [-.23, -.06], Sobel's $z = -4.06$, $p < .001$. The direct effect of trolling on culpability judgment (*c'*) was reduced but remained significant when controlling for positive affect, $t(220) = -2.27$, $p = .02$, which indicates partial mediation.

Discussion

The current study replicates the findings of Trémolière and Djeriouat (2016) linking sadistic personality to denial of responsibility for harm. Our study extends those results to delineate the psychological processes that accompany and support sadistic behaviors such as online trolling. We found that both trolls and sadists (i.e., high scorers on self-report measures of trolling and everyday sadism) tended to minimize the harm caused by an aggressive act, as compared to non-trolls and non-sadists (i.e., low-scorers on these measures). In addition, we found that trolls and sadists reacted more positively (i.e., with joy and happiness) while reading harmful scenarios, as compared to non-trolls and non-sadists. It is noteworthy that the sadism associations were uniquely significant and could not be explained by broader forms of antisociality (i.e., trait aggression, callous-unemotionality, and psychopathy scores). Finally, our follow-up analyses indicated that positive affect mediated the effects of sadism and trolling on judgments of culpability for harm. In both cases, the associations with perpetrator culpability were statistically explained by positive affect.

General Discussion

In recent years, the global Internet community has witnessed the (rather grotesque) birth of the online style known as trolling. Its proponents are brazen and confrontational; they vigorously provoke others while shrouded in online anonymity. The unfettered nature of this new social playground has revealed, if not encouraged, a sadistic subgroup of players. No longer limited to anecdotal reports, evidence from empirical research is fast accumulating that online

trolls enjoy being cruel to others: Indeed, our earlier survey research was the first to demonstrate a reliable link between trolling tendencies and the sadistic personality (Buckels et al., 2014).

Less understood is the psychological process whereby everyday sadists act on their appetites. Those motivated by the rewards of cruelty may be deterred by concomitant anxiety, regret, and felt responsibility for suffering they cause. An effective rationalization mechanism may be required to continue perpetuating such sadistic behavior as online trolling while maintaining a positive self-image.

The present research elucidated the rationalization process in two ways. In Study 1, we showed that trolls minimize others' suffering at the same time they report pleasure from that suffering. We found a parallel pattern for sadists and trolls to underestimate pain intensity, relative to non-sadists and non-trolls. In Study 2, we replicated a previously documented effect of sadism on moral judgment (Trémolière & Djeriouat, 2016), and found evidence that online trolls have similar moral deficiencies. In particular, these individuals minimize perpetrator culpability in judgments of harmful behavior. The fact that they do so in judging others' misbehavior strongly suggests that they do so in judging their own behavior.

We went further to show that sadism and trolling evidenced parallel mediation patterns: Specifically, the relations between sadism/trolling and judgments of harm were statistically explained by positive affect. Sadism emerged as an independent predictor when controlling for broader antisociality. Taken together, our results confirm that online trolls are dispositionally sensitive to the rewards afforded by interpersonal cruelty and humiliation of others. We conclude that trolling is an instance of everyday sadism. Trolling is fueled by sadistic pleasure and unleashed by rationalization.

Limitations and Recommendations

Although our results strongly confirm the sadism hypothesis of online trolling, we acknowledge that such a conclusion likely represents an oversimplification of a complex behavior. Not all trolls are sadists. There are undoubtedly many factors underlying trolling behavior—including some that are social-situational (e.g., Cheng, Bernstein, Danescu-Niculescu-Mizil, & Leskovec, 2017; Suler, 2004), and others that are dispositional in nature. Users may engage in trolling behaviors without explicitly identifying as a "troll." There may even be different types of trolls (Fichman & Sanfilippo, 2016), both within and across various social media platforms. These distinctions should be examined in future empirical research.

It is important to note that we did not specify a definition of online trolling in our questionnaires; we relied instead on participants' own interpretations of that term. Thus our results may not generalize to all trolls (only, perhaps, the more stereotypical varieties). Yet even with this possible source of "noise" in the data, we documented (a) clear convergence among various indicators of trolling (including rated enjoyment, perceived skill, and scores on our iTroll questionnaire), (b) an associated personality profile that provides a face-valid explanation for typical trolling behaviors (i.e., sadistic tendencies), and (c) responses to suffering and moral violations that are consistent with a sadistic personality profile.

We anticipate that the iTroll questionnaire will be a useful assessment tool for research in this area. In our unpublished datasets with university students, we observed acceptable test-retest correlations for iTroll scores (i.e., *r*'s > .60 for a 4 month interval). There are, however, limitations of self-reports that may be problematic when studying deceptive individuals, such as trolls. In multiple unpublished datasets, we found significant negative associations between trolling and self-reported honesty. These low honesty ratings may reflect actual dishonesty among trolls, or trolls may be overclaiming dishonesty (when they are in fact quite frank in their

responses). Alternatively, perhaps the association is driven by non-trolls' tendencies to exaggerate their honesty levels. While the meaning of the association may be unclear, we nevertheless offer a cautionary note on this issue. Whenever possible, researchers should use behavioral indicators of antisocial tendencies, and avoid relying exclusively on self-reports. Peer reports may provide invaluable validity data. Finally, it is necessary to empirically distinguish trolling from cyberbullying and other forms of cyber-deviance linked to dark traits (e.g., Smoker & March, 2017; van Geel et al., 2017). The development of validated measures such as the iTroll questionnaire should assist with that goal.

New Directions

Although preliminary, the results of Study 1 allude to an interesting paradox in the psychology of everyday sadism. By minimizing others' suffering, everyday sadists effectively diminish their hedonic returns from that suffering. This research is the first to document such a pattern among everyday sadists. It represents a clear departure from the psychology of disordered sadism, as severe sexual sadists show an opposite tendency to exaggerate others' suffering (Harenski, Thornton, Harenski, Decety, & Kiehl, 2012).[7] Our results may simply reflect a response bias generated by the nature of the rating task. Yet it is possible that our results reflect self-deceptive and/or perceptual biases that facilitate enjoyment of suffering among everyday sadists—who, unlike their disordered counterparts, presumably require some "cognitive gymnastics" to rationalize pleasure from cruelty (see Russell, & King, 2016, 2017).

One potential limitation is our interpretation of the mediation analyses. Although suggestive, we cannot infer strong causation because we used concurrent measures and collected explicit enjoyment ratings. These issues may be addressed by alternative methodologies: For example, we are currently collecting evidence of automatic pleasure responses to violent stimuli,

[7] Note, however, that their key results were based on 15 participants.

namely, facial EMG measures (Dufner & Paulhus, 2017). Social psychological methods involving implicit affect and cognition appear promising for research on everyday sadism (Mededović, 2017; Reidy, Zeichner, & Seibert, 2011). To date, some of the most compelling evidence for everyday sadism involves pleasure-driven aggression evoked in the lab (Buckels et al., 2013; Chester, 2017; Chester & DeWall, 2017a, 2017b). Chronic appetitive aggression may be contrasted with instrumentally-motivated aggression (Jones & Paulhus, 2010) and masochistic self-harm (Lämmle, Oedl, & Ziegler, 2014), which are linked to other personality tendencies.

A second, similar paradox emerged in Study 2. We found that sadism predicted lenient judgments of perpetrators of intentional harm. If not for previous findings, one might actually predict that sadists would be especially punitive (after all, is harsh and unreasonable punishment not a form of cruelty?). Indeed, there was a trend for sadists to judge perpetrators of *accidental* harm as culpable for their actions, and an overall pattern of negative association between culpability scores for accidental harm and intentional harm. Perhaps sadists only enjoy hurting incompetent or interpersonally weak targets—and not other predatory individuals (cf. Fromm, 1973). Pleasure from morally-righteous punishment may be better explained by schadenfreude (James, Kavanagh, Jonason, Chonody, & Scrutton, 2014; Porter, Bhanwer, Woodworth, & Black, 2014), revenge (e.g., Gollwitzer, Meder, & Schmitt, 2011), or prosocial motives (e.g., empathy for victims), than by sadistic tendencies.

The present research has focused on sadistic personality because it entails the appetite for cruelty that we believe embodies online trolling (it makes "lulz" possible). Empirically, however, there is evidence that psychopathy also predicts online trolling. This state of affairs is partly due to empirical overlap between sadism and psychopathy, as assessed by the available measures (e.g., see Buckels et al., 2013; Plouffe, Saklofske, & Smith, 2017; Međedović, &

Petrović, 2015). It is critical to theoretically and empirically distinguish these two constructs in future research. To this end, our lab is refining a Short Dark Tetrad questionnaire (SD4; Jones et al., in prep.), which may help resolve measurement issues in the dark personality space. Even with imperfect assessment tools, meaningful differences emerge with methodology as varied as behavioral (Buckels et al. 2013; Carre & Jones, 2016; Jones & Paulhus, 2017; Pfattheicher, Keller, & Knezevic, 2017; Rogers, Le, Buckels, Kim, & Biesanz, 2018), physiological (Dane, Jonason, & McCaffrey, 2017), and self-report (Burris & Leitch, 2017; Birkás, Gács, & Csathó, 2016; Chabrol, Melioli, Van Leeuwen, Rodgers, & Goutaudier, 2015; Duspara & Greitemeyer, 2017; Jonason et al., 2017; Neria, Vizcaino, & Jones, 2016; Plouffe, Saklofske, & Smith, 2017; Sagioglou & Greitemeyer, 2016). Scholarly appreciation and empirical attention to these differences will advance our understanding of the many varieties of dispositional malevolence in everyday life. Online trolling, we suspect, is but one everyday manifestation of sadistic tendencies. Laid bare by unrestricted Internet communication methods, this new cyber-behavior is driven by a disturbingly-ordinary appetite for cruelty, and disinhibited by effective psychological defenses.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Preparation of this manuscript was supported by an insight grant (435-2015-0417) and doctoral scholarship (767-2012-2544) awarded by the Social Sciences and Humanities Research Council of Canada.

References

Burris, C. T., Leitch, R. (2018). Harmful fun: Pranks and sadistic motivation. *Motivation and Emotion, 42*, 90-102.

Birkás, B., Gács, B., & Csathó, Á. (2016). Keep calm and don't worry: Different Dark Triad traits predict distinct coping preferences. *Personality and Individual Differences*, *88*, 134-138.

Buckels, E. E., Jones, D. N., & Paulhus, D. L. (2013). Behavioral confirmation of everyday sadism. *Psychological Science*, *24*, 2201-2209.

Buckels, E. E., & Paulhus, D. L. (2013). *Comprehensive Assessment of Sadistic Tendencies (CAST)*. Unpublished measure, University of British Columbia.

Buckels, E. E., Trapnell, P. D., & Paulhus, D. L. (2014). Trolls just want to have fun. *Personality and Individual Differences, 67*, 97-102.

Buss, A. H., & Perry, M. (1992). The aggression questionnaire. *Journal of Personality and Social Psychology*, *63*, 452-459.

Carre, J. R., & Jones, D. N. (2016). The impact of social support and coercion salience on Dark Triad decision making. *Personality and Individual Differences*, *94*, 92-95.

Chabrol, H., Melioli, T., Van Leeuwen, N., Rodgers, R., & Goutaudier, N. (2015). The Dark Tetrad: Identifying personality profiles in high-school students. *Personality and Individual Differences*, *83*, 97-101.

Chabrol, H., Van Leeuwen, N., Rodgers, R., & Séjourné, N. (2009). Contributions of psychopathic, narcissistic, Machiavellian, and sadistic personality traits to juvenile delinquency. *Personality and Individual Differences*, *47*, 734-739.

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2017). Anyone Can Become a Troll. *Proceedings of the 2017 ACM Conference on Computer Supported*

Cooperative Work and Social Computing - CSCW '17. doi:10.1145/2998181.2998213

Cheng, C., Lin, C., Liu, H., Hsu, Y., Lim, K., Hung, D., & Decety, J. (2007). Expertise modulated the perception of pain in others. *Current Biology*, *17*, 1708-1713.

Chester, D. S. (2017). The role of positive affect in aggression. *Current Directions in Psychological Science*, *26*, 366-370.

Chester, D. S. & DeWall, C. N. (2017a). Combating the sting of rejection with the pleasure of revenge: A new look at how emotion shapes aggression. *Journal of Personality and Social Psychology*, *112*, 413-430.

Chester, D.S. & DeWall, C.N. (2017b). Personality correlates of revenge-seeking: Multidimensional links to physical aggression, impulsivity, and aggressive pleasure. *Aggressive Behavior*. Advance online publication. doi: 10.1002/ab.21746

Coles, B. A., & West, M. (2016). Trolling the trolls: Online forum users' constructions of the nature and properties of trolling. *Computers in Human Behavior*, *60*, 233-244.

Craig K. D., Versloot, J., Goubert, L., Vervoort, T., Crombez, G. (2010). Perceiving pain in others: automatic and controlled mechanisms. *Journal of Pain*, *11*, 101-108.

Craker, N., & March, E. (2016). The dark side of Facebook®: The Dark Tetrad, negative social potency, and trolling behaviours. *Personality and Individual Differences, 102*, 79-84.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*, 353-380.

Dane, L. K., Jonason, P. K., & McCaffrey, M. (2017). Physiological tests of the cheater hypothesis for the Dark Triad traits: Testosterone, cortisol, and a social stressor. *Personality and Individual Differences*.

Dewey, C. (2016, March 23). Meet Tay, the creepy-realistic robot who talks just like a teen. *The*

Washington Post. Retrieved 19 Nov 2016 from https://www.washingtonpost.com/news/the-intersect/wp/2016/03/23/meet-tay-the-creepy-realistic-robot-who-talks-just-like-a-teen/

Dufner, M., & Paulhus, D. L. (2017). *Distinctions among the Dark Tetrad using facial electromyography*. Unpublished manuscript.

Duspara, B., & Greitemeyer, T. (2017). The impact of dark tetrad traits on political orientation and extremism: an analysis in the course of a presidential election. *Heliyon*, 3(10), e00425.

Dynel, M. (2016). "Trolling is not stupid": Internet trolling as the art of deception serving entertainment. *Intercultural Pragmatics*, *13*, 353-381.

Fichman, P., & Sanfilippo, M. R. (2016). *Online trolling and its perpetrators: Under the cyberbridge*. Rowman & Littlefield.

Foulkes, L., McCrory, E. J., Neumann, C. S., & Viding, E. (2014). Inverted social reward: Associations between psychopathic traits and self-report and experimental measures of social reward. *PloS one*, *9*(8), e106000.

Frick, P. J., & Ray, J. V. (2015). Evaluating callous-unemotional traits as a personality construct. *Journal of Personality*, *83*, 710-722.

Fromm, E. (1973). *The anatomy of human destructiveness*. NY: Holt, Rinehart and Winston.

Gammon, A. (2014). *Over a quarter of Americans have made malicious online comments*. Retrieved September 7, 2015, from https://today.yougov.com/news/2014/10/20/over-quarter-americans-admit-malicious-online-comm/

Gollwitzer, M., Meder, M., & Schmitt, M. (2011). What gives victims satisfaction when they seek revenge? *European Journal of Social Psychology*, *41*, 364-374.

Greitemeyer, T. (2015). Everyday sadism predicts violent video game preferences. *Personality and Individual Differences*, *75*, 19-23.

Greitemeyer, T., & Sagioglou, C. (2017). The longitudinal relationship between everyday sadism and the amount of violent video game play. *Personality and Individual Differences*, *104*, 238-242.

Hardaker, C. (2010). Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *Journal of Politeness Research, 6*, 215-242.

Harenski, C. L., Thornton, D. M., Harenski, K. A., Decety, J., Kiehl, K. A. (2012). Increased frontotemporal activation during pain observation in sexual sadism: Preliminary findings. *Archives of General Psychiatry*, *69*, 283-92.

James, S., Kavanagh, P. S., Jonason, P. K., Chonody, J. M., & Scrutton, H. E. (2014). The Dark Triad, schadenfreude, and sensational interests: Dark personalities, dark emotions, and dark behaviors. *Personality and Individual Differences*, *68*, 211-216.

Jonason, P. K., Foster, J. D., Egorova, M. S., Parshikova, O., Csathó, Á., Oshio, A., & Gouveia, V. V. (2017). The Dark Triad traits from a life history perspective in six countries. *Frontiers in Psychology*, *8*, 1476.

Jones, D. N., Buckels, E. E., & Paulhus, D. L. (2017). *A brief measure of the Dark Tetrad (SD4).* Unpublished manuscript, University of British Columbia.

Jones, D. N., & Paulhus, D. L. (2010). Different provocations trigger aggression in narcissists and psychopaths. *Social Psychological and Personality Science*, *1*, 12-18.

Jones, D. N., & Paulhus, D. L. (2014). Introducing the Short Dark Triad (SD3): A brief measure of dark personality traits. *Assessment*, *21*, 28–41.

Jones, D. N., & Paulhus, D. L. (2017). Duplicity among the Dark Triad: Three faces of deceit.

Journal of Personality and Social Psychology, *113*, 329-342.

Lämmle, L., Oedl, C., & Ziegler, M. (2014). Don't threaten me and my dark side or even self-harm won't stop me from hurting you. *Personality and Individual Differences*, *67*, 87-91.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). *International affective picture system (IAPS): Affective ratings of pictures and instruction manual*. Technical Report A-8. University of Florida, Gainesville, FL.

Leone, M. (2017). *The art of trolling*. Unpublished manuscript, University of Turin.

Maltby, J., Day, L., Hatcher, R. M., Tazzyman, S., Flowe, H. D., Palmer, E. J., . . . Cutts, K. (2016). Implicit theories of online trolling: Evidence that attention-seeking conceptions are associated with increased psychological resilience. *British Journal of Psychology, 107*, 448-466.

March, E., Grieve, R., Marrington, J., & Jonason, P. K. (2017). Trolling on Tinder® (and other dating apps): Examining the role of the Dark Tetrad and impulsivity. *Personality and Individual Differences*, *110*, 139-143.

Marshall, J., Watts, A. L., & Lilienfeld, S. O. (2018). Do psychopathic individuals possess a misaligned moral compass? A meta-analytic examination of psychopathy's relations with moral judgment. *Personality Disorders: Theory, Research, and Treatment*, *19*, 40-50.

Međedović, J., (2017). Aberrations in emotional processing of violence-dependent stimuli are the core features of sadism. *Motivation and Emotion*, 41, 273–283.

Međedović, J., & Petrović, B. (2015). The dark tetrad: Structural properties and location in the personality space. *Journal of Individual Differences*, *36*, 228–236.

Nell, V. (2006). Cruelty's rewards: The gratifications of perpetrators and spectators. *Behavioral and Brain Sciences*, *29*, 211-224.

Neria, A. L., Vizcaino, M., & Jones, D. N. (2016). Approach/avoidance tendencies in dark personalities. *Personality and Individual Differences*, *101*, 264-269.

Ohlheiser, A. (2016, March 25). Trolls turned Tay, Microsoft's fun millennial AI bot, into a genocidal maniac. *The Washington Post*. Retrieved 19 Nov 2016 from https://www.washingtonpost.com/news/the-intersect/wp/2016/03/24/the-internet-turned-tay-microsofts-fun-millennial-ai-bot-into-a-genocidal-maniac/

O'Meara, A., Davies, J., & Hammond, S. (2011). The psychometric properties and utility of the short sadistic impulse scale (SSIS). *Psychological Assessment*, *23*, 523–531.

Pfattheicher, S., Keller, J., & Knezevic, G. (2017). Sadism, the intuitive system, and antisocial punishment in the public goods game. *Personality and Social Psychology Bulletin*, *43*, 337-346.

Paulhus, D. L. (2014). Toward a taxonomy of dark personalities. *Current Directions in Psychological Science, 23,* 421-426.

Paulhus, D. L., & Dutton, D. G. (2016). Everyday sadism. In V. Zeigler-Hill & D. K. Marcus (Eds.), *The dark side of personality: Science and practice in social, personality, and clinical psychology* (pp. 109-120). Washington, D.C.: American Psychological Association.

Paulhus, D. L., & Jones, D. N. (2015). Measures of dark personalities. In G. J. Boyle, D. H. Saklofske, & G. Matthews (Eds.), *Measures of personality and social psychological constructs* (pp. 562-594). San Diego: Academic Press.

Paulhus, D. L., & Williams, K. M. (2002). The Dark Triad of personality: Narcissism, Machiavellianism and psychopathy. *Journal of Research in Personality, 36*, 556-563.

Phillips, W. (2015). *This is why we can't have nice things: Mapping the relationship between*

online trolling and mainstream culture. MIT Press.

Plouffe, R. A., Saklofske, D. H., & Smith, M. M. (2017). The assessment of sadistic personality: Preliminary psychometric evidence for a new measure. *Personality and Individual Differences*, *104*, 166-171.

Porter, S., Bhanwer, A., Woodworth, M., & Black, P. J. (2014). Soldiers of misfortune: An examination of the Dark Triad and the experience of schadenfreude. *Personality and Individual Differences*, *67*, 64-68.

Ray, J. V., Frick, P. J., Thornton, L. C., Steinberg, L., & Cauffman, E. (2016). Positive and negative item wording and its influence on the assessment of callous-unemotional traits. *Psychological Assessment*, *28*, 394-404.

Reidy, D. E., Shelley-Tremblay, J. F., & Lilienfeld, S. O. (2011). Psychopathy, reactive aggression, and precarious proclamations: A review of behavioral, cognitive, and biological research. *Aggression and Violent Behavior*, *16*, 512-524.

Rodriguez, A. (2016, March 24). Microsoft's AI millennial chatbot became a racist jerk after less than a day on Twitter. *Quartz*. Retrieved 19 Nov 2016 from http://qz.com/646825/microsofts-ai-millennial-chatbot-became-a-racist-jerk-after-less-than-a-day-on-twitter/

Rogers, K. H., Le, M. T., Buckels, E. E., Kim, M., & Biesanz, J. C. (2018). Dispositional malevolence and impression formation: Dark Tetrad associations with accuracy and positivity in first impressions. *Journal of Personality.* Advance online publication. doi:10.1111/jopy.12374

Russell, T. D., & King, A. R. (2017). Mean girls: PID-5 personality traits and everyday sadism predict hostile femininity. *Personality and Individual Differences*, *104*, 252-257.

Russell, T. D., & King, A. R. (2016). Anxious, hostile, and sadistic: Maternal attachment and everyday sadism predict hostile masculine beliefs and male sexual violence. *Personality and Individual Differences*, *99*, 340-345.

Sest, N., & March, E. (2017). Constructing the cyber-troll: Psychopathy, sadism, and empathy. *Personality and Individual Differences*, *119*, 69-72.

Shachaf, P., & Hara, N. (2010). Beyond vandalism: Wikipedia trolls. *Journal of Information Science*, *36*, 357–370.

Schumpe, B. M., & Lafreniere, M. A. K. (2016). Malicious joy: Sadism moderates the relationship between schadenfreude and the severity of others' misfortune. *Personality and Individual Differences*, *94*, 32-37.

Smoker, M., & March, E. (2017). Predicting perpetration of intimate partner cyberstalking: gender and the Dark Tetrad. *Computers in Human Behavior*, *72*, 390-396.

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & Behavior*, *7*, 321-326.

Trémolière, B., & Djeriouat, H. (2016). The sadistic trait predicts minimization of intention and causal responsibility in moral judgment. *Cognition*, *146*, 158-171.

van Geel, M., Goemans, A., Toprak, F., & Vedder, P. (2017). Which personality traits are related to traditional bullying and cyberbullying? A study with the Big Five, Dark Triad and sadism. *Personality and Individual Differences*, *106*, 231-235.

Webster, G. D., DeWall, C. N., Pond, R. S., Deckman, T., Jonason, P. K., Le, B. M., ... & Smith, C. V. (2014). The brief aggression questionnaire: Psychometric and behavioral evidence for an efficient measure of trait aggression. *Aggressive Behavior*, *40*, 120-139.

Table 1

Trait Associations of Online Trolling in Studies 1 and 2

	Online Trolling (GAIT)	
Study 1 – Trait Measures	*r*	*β*
Sadism (CAST)	.71***	.56***
Machiavellianism	.32***	-.04
Narcissism	.26***	.03
Psychopathy	.63***	.21**
Study 2 – Trait Measures	Online Trolling (iTroll)	
Sadism (SSIS)	.44***	.24**
Aggression	.31***	-.02
Callous-Unemotionality	.36***	.11
Psychopathy	.43***	.23*

Note. Study 1 (Valid $N = 304$), Study 2 (Sample B; Valid $N = 223$). Tabled values are bivariate correlations (r's) and standardized regression weights when controlling for scores the other trait measures (β's). * $p < .05$. ** $p < .01$. *** $p < .001$.

Table 2

Associations with Perceived Pain Intensity and Pleasure from Pain

Measure	Perceived Pain Intensity	Pleasure from Pain
Sadism	-.27*** [-.46***]	.46*** [.59***]
Trolling	-.26*** [-.45***]	.40*** [.53***]
Machiavellianism	-.14* [-.25*]	.23*** [.31***]
Narcissism	-.09 [-.17]	.13* [.18*]
Psychopathy	-.23*** [-.40***]	.42*** [.56***]

Note. Valid $N = 304$. Tabled values are bivariate correlations (r's). Values in square parentheses are corrected for attenuation.

* $p < .05$; ** $p < .01$; *** $p < .001$.

Table 3

Correlations of iTroll with Validity Measures and Sadistic Personality

Measure	1	2	3	4
1. iTroll				
2. Sadism	.46***			
3. Trolling enjoyment	.51***	.38***		
4. Trolling frequency	.33***	.30***	.26***	
5. Trolling skill	.70***	.58***	.57***	.36***

Note. * $p < .05$; ** $p < .01$; *** $p < .001$.

Table 4

Predictors of Culpability Judgments within the Intentional, Attempted, and Accidental Harm Conditions

	Moral Culpability					
Measure	Intentional Harm		Attempted Harm		Accidental Harm	
Trolling	-.28***	(-.20**)	-.23***	(-.13)	.03	(-.04)
Sadism	-.28***	(-.20*)	-.30***	(-.22*)	.14*	(.13)
Aggression	-.17*	(.00)	-.26***	(-.17*)	.10	(.03)
Callous-Unemotionality	-.12	(.11)	-.17*	(.03)	.09	(.02)
Psychopathy	-.23**	(-.08)	-.22*	(.08)	.10	(.01)

Note. $N = 236$. Tabled values without parentheses are bivariate correlations (*r*'s). Values in parentheses are standardized regression weights (β's) controlling for the other predictor variables. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

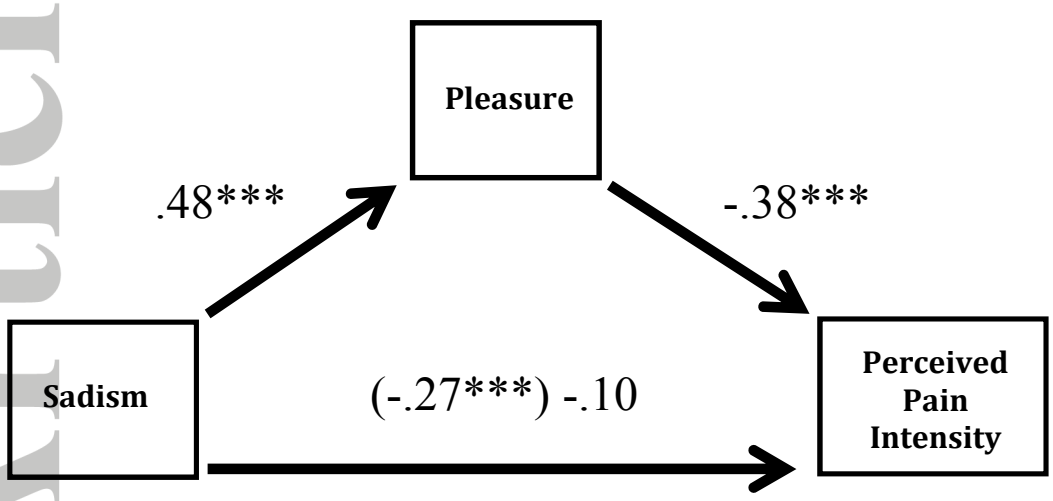
Table 5

Predictors of Positive Affect within the Intentional, Attempted, and Accidental Harm Conditions

	Positive Affect					
Measure	Intentional Harm		Attempted Harm		Accidental Harm	
Trolling	.33***	(.22**)	.31***	(.24**)	.39***	(.28***)
Sadism	.38***	(.34***)	.32***	(.32***)	.38***	(.27**)
Aggression	.15*	(-.07)	.11	(-.07)	.17*	(-.09)
Callous-Unemotionality	.25***	(.06)	.08	(-.17*)	.25***	(.04)
Psychopathy	.22**	(-.09)	.20**	(.03)	.28***	(.02)

Note. $N = 236$. Tabled values without parentheses are bivariate correlations (r's). Values in parentheses are standardized regression weights (β's) controlling for the other predictor variables. *$p < .05$; **$p < .01$; ***$p < .001$.

(Model A)



(Model B)

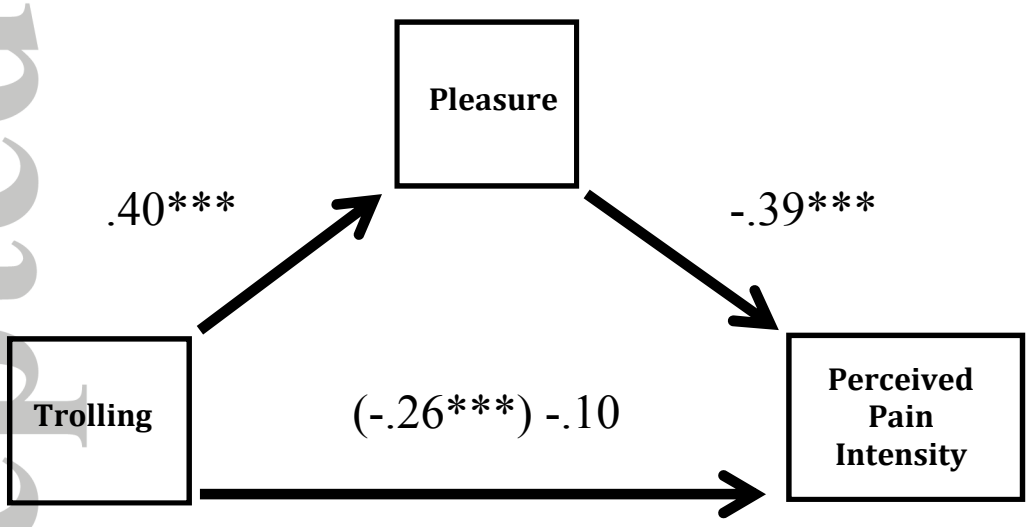


Figure 1. Self-reported pleasure from others' pain as a mediator of the relationships between sadism and perceived pain intensity (Model A), and trolling and perceived pain intensity (Model B) in Study 1. Path coefficients are standardized regression coefficients (β's). Those in parentheses are from the main effects model prior to adding the mediating term. * $p < .05$. *** $p < .001$.

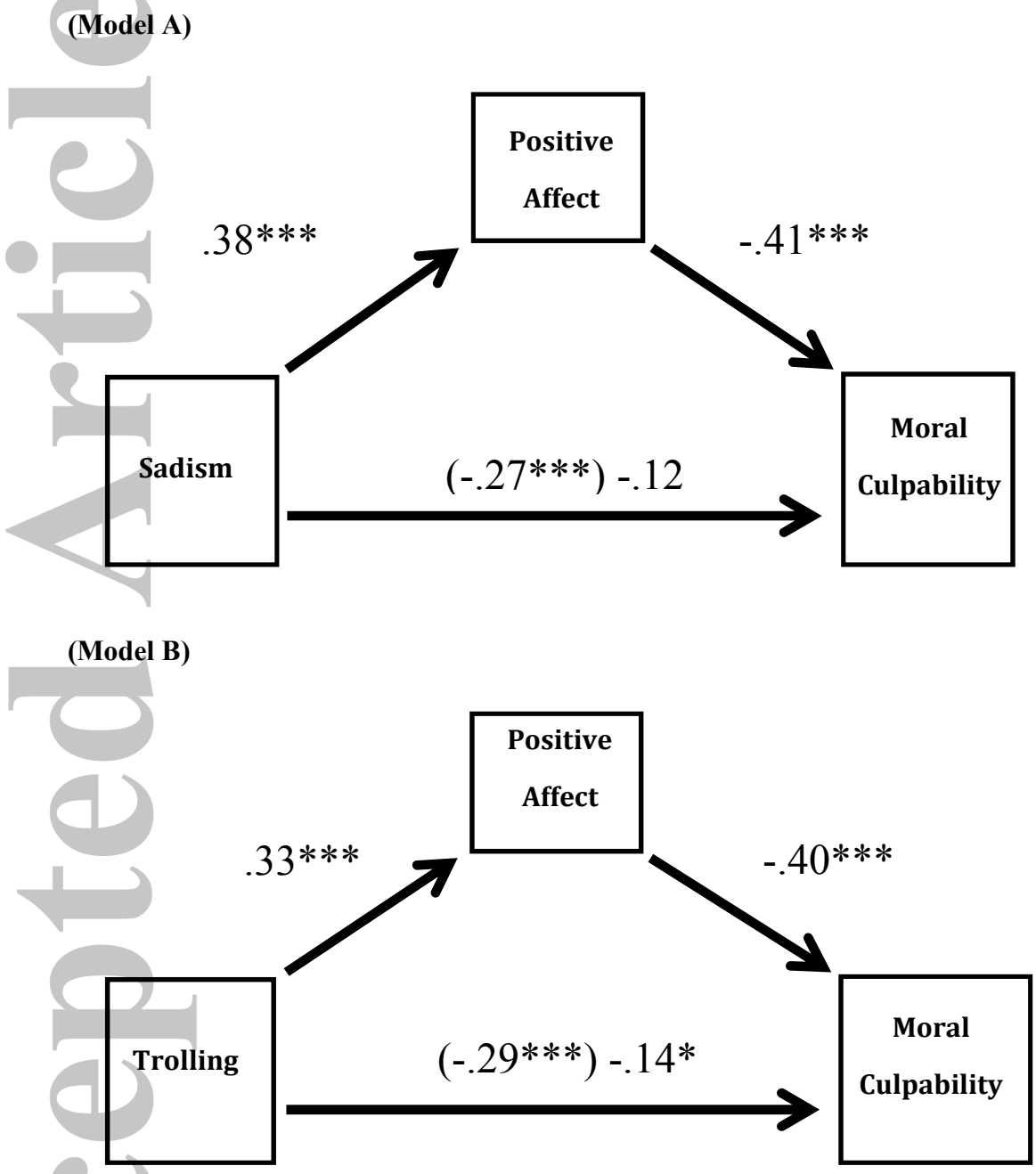


Figure 2. Positive affect as a mediator of the relationship between sadism and culpability judgment (Model A) and trolling and culpability judgment (Model B) in the intentional harm condition of Study 2. Path coefficients are standardized regression coefficients (β's). Those in parentheses are from the main effects model prior to adding the mediating term. * $p < .05$. ** $p < .01$. *** $p < .001$.