Social (media) vigilantes: Effects of social vigilantism and anonymity on online confrontations of prejudice

by

Tiffany Jo Lawless

B.S., Cornell College, 2015
M.S., Kansas State University, 2019

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Psychological Sciences
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2023

Abstract

Research has focused on how anonymity affects perceptions of prejudice online (e.g., Lawless & Saucier, submitted a), but it is possible individual differences like Social Vigilantism (SV), the tendency to impress one's beliefs onto others, also affect these perceptions (Saucier & Webster, 2010). In Study 1, SV is measured and participants see mock prejudiced posts in a within-groups 2(anonymous/identifiable) x 2(including/not including comments confronting posts) design and rate perceptions of posts (e.g., *The person who posted this is racist*). Study 2 uses the same methods as Study 1 but asks participants how they would interact with posts. It is shown that SV is associated with more confrontation of prejudice because SV is associated with counterarguing. It is also demonstrated that anonymity of platform affects the results such that posts on identifiable platforms receive more interactions because they are seen as more honest.

Social (media) vigilantes: Effects of social vigilantism and anonymity on online confrontations of prejudice

by

Tiffany Jo Lawless

B.S., Cornell College, 2015
M.S., Kansas State University, 2019

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Psychological Sciences
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2023

Approved by:

Major Professor
Donald A. Saucier

Copyright

Abstract

Research has focused on how anonymity affects perceptions of prejudice online (e.g., Lawless & Saucier, submitted a), but it is possible individual differences like Social Vigilantism (SV), the tendency to impress one's beliefs onto others, also affect these perceptions (Saucier & Webster, 2010). In Study 1, SV is measured and participants see mock prejudiced posts in a within-groups 2(anonymous/identifiable) x 2(including/not including comments confronting posts) design and rate perceptions of posts (e.g., *The person who posted this is racist*). Study 2 uses the same methods as Study 1 but asks participants how they would interact with posts. It is shown that SV is associated with more confrontation of prejudice because SV is associated with counterarguing. It is also demonstrated that anonymity of platform affects the results such that posts on identifiable platforms receive more interactions because they are seen as more honest.

Table of Contents

List of Tables

Acknowledgements

Thank you to all my family, friends, mentors, and colleagues who supported me during the writing of this dissertation.

Dedication

To you, because you actually read the dedication of a dissertation.

Chapter 1 - Anonymity and Social (Media) Vigilantes

While social media has upgraded communication and has allowed individuals to more easily share stories and productive discussions, it has also created new paths for the expression of hate and prejudice. The tendency of anonymity on the internet to negatively influence behavior is widely recognized and is generally accepted by both the scientific and lay communities (e.g., Reader, 2012). Anonymous posts online tend to be more aggressive than identifiable posts (Moore, Nakano, Enomoto, & Suda, 2012), and people who believe they are anonymous exhibit cyber aggression more often (Wright, 2013). When people behave anonymously, they are less likely to display altruism (Locey & Rachlin, 2015), and more likely to display harmful behavior (Nogami & Takai, 2008). While past studies have focused on how anonymity of both the perpetrator and target affects third-party perceptions of racial prejudice online (Lawless & Saucier, under review a; under review b; in preparation), little is known about how individual differences within the third-party affect perceptions of such online interactions. It is possible that individual differences, including social vigilantism and personal endorsement of prejudicial beliefs, affect how third parties perceive and would choose to interact with prejudicial statements online.

Anonymity is the degree to which the recipient or spectator of messages perceives the message's source as unidentified or unknown (Scott, 1998). Anonymity has become a fundamental fact of online communication, with common instances such as anonymous usernames and "throwaway" accounts as well as rarer occurrences such as one person holding and automatically communicating from hundreds of email addresses and social media accounts at one time. Whereas talking anonymously online with others had relatively benevolent origins in cooperative text-based videogames and fandom and hobby message boards, it has now become a

method used widely by both political and anarchistic ‘hacktivism’ groups such as the aptly named Anonymous as well as superficially harmless but arguably sinister “trolls” and cyberbullies (Crawford, 2009; Wang, Wang, Wang, Nika, Zheng, & Zhao, 2014). Removing the usual perceptions of strong individualities and long-term social associations, largely anonymous social media platforms, such as Reddit, Whisper, Discord, and 4chan encourage communication between strangers and allow users to express themselves virtually without fear of real-life penalties (Crawford, 2009; Wang, et al., 2014).

As prospects for online communication have increased, there has also been a rise in new social problems, such as cyberbullying. This is particularly true among adolescents, but cyber aggression has also increased in adult populations (Schnurr, Mahatmya, & Basche III, 2013; Sugarman & Willoughby, 2013). Hinduja and Patchin (2008) define cyberbullying as a deliberate, recurring, and emotionally harmful action using a computer, mobile phone, or other electronic device. The rapid growth in technology presents many paths for cyberbullies to participate in damaging behavior, and it also allows cyberbullies to remain concealed from victims and law enforcement if they choose an anonymous platform. A cyberbully can quickly use a digital device to post or direct a cruel message to a large group of people while keeping their identity unobserved. By using Facebook, Instagram, or Twitter, someone who cyberbullies can post a hurtful message about their targets, and in a very short timeframe, it will show up in the feeds of the target’s real-life friends and acquaintances. A malicious message can be seen in seconds by a plethora of online users. Even after it has been deleted, a message can be re-discovered because nothing can reliably be completely removed from the internet. One way cyberbullying is happening is on social networking sites such as Facebook, Instagram, Snap Chat, and Twitter. Using social media platforms to engage in cyberbullying is almost unanimous

among teens; 95% of teens on social media have witnessed malicious or harsh behavior on social media sites (Lenhart, Madden, Smith, Purcell, Zickuhr, & Rainie, 2011).

Early research on cyberbullying created personality profiles of those involved in cyberbullying as targets and/or bullies. Troublingly, one study found people with special needs, unusual academic abilities, poor social skills, odd or undesirable physical appearances, physical and mental disabilities, unfashionable clothing, those of a minority ethnicity, and other marginalized groups were often targeted (Cassidy, Jackson, & Brown, 2009). Thus, seeing racist and prejudicial attacks online is an unfortunately common situation. Such pervasive exposure may numb observers to this kind of communication, and this may result from and help create an online culture with distinct social and moral norms from face-to-face culture.

Runions (2015) suggests that technological communication modifies views toward victims of cyberviolence; third-party observers experience moral detachment from their usual values. The tonal and social uncertainty created by virtual communication alters perceptions of culpability and empathy, with third-party observers justifying their reactions (or lack thereof) by perceiving the victim as somehow to blame for the attack (Runions, 2015).

Additionally, third-party observers involve themselves in instances of cyberbullying more often as compared to face-to-face bullying, and they directly change the bullying experience through their involvement (Anderson, Bresnahan, & Musatics, 2014; Barlinska, Szuster, & Winiewski, 2013; Runions, 2015). The process of communicating through technology may modify the social context or way that third-party observers interact with victims of bullying (Calvete, Orue, Estévez, Villardón, & Padilla, 2010). Third-party observers may respond to an atmosphere that models aggression by adding to the disparagement of the victims. Another unique characteristic is that a victim of cyberbullying may be uninformed of who the bully is.

One study found 48% of those bullied did not know who had bullied them because the bully had kept an anonymous username or had used an anonymous platform (Kowalski, & Limber, 2007). One study found that when cyberbullying was anonymous, there was little emotional effect on the victim (Reeckman & Cannard, 2009). However, other studies have been unable to reproduce this outcome (e.g., Dilmac, 2009; Price & Dalgleish, 2010). Additionally, many results have shown a link between aggression and attention-seeking and committing cyberbullying behaviors. (Harman, Hansen, Cochran, & Lindsey, 2005; Li, 2005; Willard, 2007). Bullies could be engaging in cyberbullying for amusement, to exercise control, or both (Cassidy et al., 2009; Reeckman & Cannard, 2009). It is possible that witnesses understand the possibility of these motivations, and therefore attribute online prejudiced speech to honesty in some circumstances and to attention-seeking or "trolling" in others (Lawless & Saucier, under review b).

Hardaker (2010) defines a troll as a person "whose real intention(s) is/are to cause disruption and/or to trigger or exacerbate conflict for the purposes of their own amusement" (pg. 237). Trolling saturates the online environment. On some level, trolls are responsible for generating and/or augmenting many popular and harmless memes. LOLcats, RickRolling, the Guy Fawkes mask, advice animals, demotivational posters, and other attention-seeking gimmicks are common entertaining trolling behaviors. However, trolls also often perform more antisocial and widely intolerable behaviors, such as promoting racist, sexist, homophobic, or other prejudicial rhetoric. Additionally, most trolls build firewalls between their online and offline selves, making their "true" intentions difficult if not impossible to determine while carefully maintaining their anonymity (Bourdieu 2001; Dahlberg, 2001; Donath, 1998). Trolling is characteristically grounded on sensationalism, spectacle, and emotional exploitation, all of which can be achieved through extreme prejudiced speech. Therefore, it is possible that people

commonly perceive online anonymous prejudiced speech targeting groups of people as attention-seeking behavior rather than as honest dissemination of personal beliefs (Lawless & Saucier, under review b). As Donath (1998) contends, the existence of trolls, or even the possibility of the existence of trolls, makes community members less likely to trust outsiders and anonymous people. This idea of trolling has saturated internet culture to the point that research has suggested that some communities (e.g., Reddit) use trolling accusations as social disincentives to prevent lying (Bergstrom, 2011). However, it is also understood that engaging in racism or sexism or homophobia does not inevitably make someone a troll, and there is therefore general ambiguity regarding the honesty of anonymous people online.

On the other hand, there is evidence that the opportunity to connect anonymously over the internet results in greater opportunities for honesty, and perhaps greater perceptions of honesty by third parties. Bargh, McKenna, and Fitzsimons (2002) found that certain social networking platforms decrease blocks to communication, which then bolsters greater self-disclosure. In an experiment where participants were randomly assigned to interact with discussion partners in an online setting or in a face-to-face setting, those who were talking online were better able to express their 'true-self' (Bargh et al., 2002). Tidwell and Walther (2002) also found that people who communicated via computer had a higher proportion of close and direct uncertainty reduction behaviors than those who met in face-to-face interactions. Additionally, Lawless and Saucier (under review b) found that anonymous prejudicial speech was perceived as more honest when directed at a named individual than when directed at a large group (e.g., Black people as a whole).

Toxic Online Disinhibition

Hatred directed at group members due to factors they cannot control, such as their race, is not new, but it has taken on a new breadth in the online world. Online hate includes actions such as the denigration, harassment, and exclusion of, as well as the encouragement of violence toward, specific groups (Hawdon, Oksanen, & Räsänen, 2017; Räsänen, Hawdon, Holkeri, Keipi, Näsi, & Oksanen, 2016). The online environment involves anonymity, invisibility, asynchronicity, textuality, and lack of face-to-face contact, and consequences are considered less likely to ensue as compared with the offline world (Suler, 2004). These conditions can encourage rude language, hatred, and threats. This tendency is also called toxic online disinhibition (Suler, 2004). Toxic online disinhibition can reduce the capacity for empathy, self-control, and recognizing social cues (Suler, 2004; Voggeser, Singh, & Göritz, 2018). Past research has revealed that higher levels of toxic online disinhibition are positively associated with cyberbullying and trolling (Görzig & Ólafsson, 2013; Udris, 2014; Voggeser et al., 2018; Wright, 2013; Wright, Harper, & Wachs, 2018). Therefore, it can be proposed that toxic online disinhibition might also lead to less self-monitoring when expressing beliefs through hateful or degrading speech online, making inappropriate attacks on marginalized groups more likely.

The online disinhibition effect is a reduction in the reservedness of behavior commonly exhibited in online environments (Joinson, 2007; Kiesler, Siegel, & McGuire, 1984; Suler 2004). Many online behaviors, particularly those completed anonymously, can be ascribed to the online disinhibition effect (Joinson, 2001; Kiesler et al., 1984). These behaviors often manifest as aggressive and/or hateful posts or comments, and these antagonistic behaviors can be ascribed to toxic online disinhibition (Suler, 2004). Posts attributable to toxic online disinhibition

characteristically include aggressive language, swearing, and disparaging or offensive names (Dyer, Green, Pitts, & Millward, 1995). Such toxic and hostile behaviors can often be found, not only within niche prejudicial blogs and instances of cyberbullying, but also in places as innocuous as online video gaming sites and the comments on YouTube videos (Chau & Xu, 2007; Huang & Chou, 2010; Moor, Heuvelman, & Verleur, 2010; Williams & Skoric, 2005). Given that anonymity is often a major factor in the development of toxic online disinhibition (Joinson, 2007; Lapidot-Lefler & Barak, 2012; Suler, 2004), and that people recognize this behavior as toxic and outside the general social norms (e.g., Reader, 2012), it is possible that people discount hateful and anonymous online activity and attribute less importance and honesty to it (e.g., Lawless & Saucier, under review b).

Social Identity Model of Deindividuation Effects (SIDE) Theory

SIDE theory suggests that technological communication modifies perceptions of oneself and others (Postmes, Spears, & Lea, 1998), contributing to conflict that encourages online prejudice. Espousing prejudiced rhetoric is an inherently social process because of the number of third-party observers that may witness the speech, particularly on the internet (Anderson et al., 2014; Barlinska et al., 2013). Via action or inaction, third-party observers can affect the severity of online prejudice for targets. Third-party observers commenting or forwarding a hateful message actively contribute to the bullying process, whereas third-party observers communicating support to the target may reduce the trauma associated with being targeted (Anderson et al., 2014). The application of SIDE to the issue of online expressions of prejudice examines how technological communication changes perceptions of identity, increasing the likelihood of acting in ways that differ from normal behavior (Postmes et al., 1998).

SIDE theory suggests that when individuals communicate through technology, a change in perception occurs (Postmes et al., 1998). Postmes and colleagues (1998) argue that the social definition participants give to a context affects how they communicate with each other through technology, and the features of that technology may in turn influence how the interaction unfolds. Postmes and Baym (2005) suggest that although the use of communication technology does not necessarily lead to uniform effects across situations, and technology does not determine interpersonal interactions en masse, it does have an influence on an individual and social level (Postmes & Baym, 2005). Namely, the features of technological communication highlight certain aspects of identity during online interactions, creating a shift in perceptions that can alter communication (Postmes et al., 1998). Technology leads certain elements of interactions to become more or less salient (Postmes et al., 1998). The noticeable effects of this salience are changes in perceptions of individual identity compared to social identity (Postmes, Spears, Sakhel, & de Groot, 2001). Postmes and Baym (2005) suggest that when communicating with others, individuals retain a sense of personal identity while maintaining a perception of social identity. SIDE postulates that the features of online communication heighten awareness of the social context or group (Postmes & Baym, 2005). Accounting for why people place importance on social identity in technological communication, Moral-Toranzo, Canto-Ortiz, and Gomez-Jacinto (2007) explain that it mollifies the need to belong and is tied to self-satisfaction. This keen cognizance of the group, or the disconfirming comments and lack of confirming comments to a victim of prejudice could influence third-party observers' perceptions when deciding whether and how they ought to respond to a hateful message. For example, individuals may participate in self-stereotyping, underpinning their features and opinions based on the dominant views of the group (Postmes & Spears, 2002). As the group identity becomes more prominent,

individuals are more likely to obey group norms. Additionally, individuals perceive some amount of anonymity when online, even if they know each other and interact in real life (Moral-Toranzo et al., 2007). Applied to third-party observers of prejudiced posts, individuals might be more likely to comment in a certain way or avoid supportive actions toward victims than they would be in real-life because the online group norms support a culture where prejudice is to be expected and perhaps even condoned.

In addition, SIDE explains that, when communicating in a technological context, sensitivities toward individual identities are reduced (Postmes & Baym, 2005). The process is called deindividuation and can explain why third-party observers might respond directly to prejudiced messages (Barlinska et al., 2013). Due to an alteration in perception that occurs when communicating through technology, third-party observers may feel a need to respond in a way that reinforces social identity (Barlinska et al., 2013). Individuals will contemplate how comments fit in with the social group viewing the message (Postmes & Baym, 2005) and lose awareness of comments being directly received by the victim, potentially leading to a lack of understanding of how the victim is harmed by such comments.

Deindividuation occurs when individuals experience reduced awareness of themselves and of others (e.g., Carr, Vitak, & McLaughlin, 2013; Postmes et al., 1998). SIDE proposes that anonymity is a crucial factor in determining how deindividuation occurs (Postmes et al., 2001). Postmes and colleagues (1998) contend that the way communication occurs through technology can lead to a modification in cognitive processing. A characteristic response to a social situation changes as anonymity reduces perceptions of personal identity and amplifies views of group identity. Postmes and Baym (2005) propose that, in social interactions, individuals have a sense of both individual and social identities, but group membership is often inflated online (Walther

& Bazarova, 2007). The asynchrony of communication and the unique capability of numerous others to reply to a social media message facilitates the perception of communicating with a group even if a message is directed to one specific member (Postmes & Baym, 2005). As a result, when a third-party observer views a prejudiced message, the communication may be thought of as reflective of the group, as opposed to personal communication. A third-party observer that experiences deindividuation would pay more attention to the social context, or the comments of others, rather than bearing in mind how the response, or lack of a sympathetic response, directly impacts the victim. Deindividuation increases with greater anonymity (Postmes et al., 2001). A lack of distinctive features modifies perceptions of the self and of others as individuals (Postmes & Baym, 2005). Exemplifying this point, Postmes and Spears (2002) found increasing anonymity by manipulating perceptions of personal identity led to greater use of gendered stereotypes, and individuals exhibited little worry that the comments would trace back to them.

Additionally, deindividuation leads to an awareness of a breakdown of traditional social barriers (e.g., Postmes et al., 2001). As a result, people may feel emboldened in their actions, behaving without inhibitions and communicating antisocially, and thus behaving contrarily to how they customarily would in a real-life context (e.g., Postmes et al., 2001). Circumstances characterized by anonymity appear to change social barriers by enabling more negative behaviors instead of encouraging equality. In the case of online prejudice, third-party observers may feel emboldened to like, share, or leave disparaging messages for victims. For example, Slonje, Smith, and Frisen (2012) applied deindividuation to explain why youths would act as cyberbullies. Slonje and colleagues (2012) suggested that deindividuated cyberbullies would feel less culpability and remorse for actions because the perception of directly bullying another

person is reduced in a cyber context. Similarly, third-party observers may feel a lack of regret about disparaging actions or inaction to support cyberbullying victims because of being deindividuated. Anonymity in cyberbullying reduces pressure and limitations when communicating with victims (Calvete et al., 2010). When a social media site alters social cues and offers a sense of protection through anonymity, internet users may feel emboldened similar to the findings of anonymity in situations of mob mentality (Calvete et al., 2010; Runions, 2015). Barlinska and colleagues (2013) suggest that the sense of deindividuation is expanded by a lack of direct criticism from victims. In other words, without victims articulating the injury they experienced, the sense of one's actions negatively affecting the victims is diminished. Thus, internet users experience reduced accountability for behaviors. Anonymity is an important aspect accounting for cyberbullying, and SIDE provides a context showing the effects of anonymity on deindividuation. Another aspect of deindividuation should be examined in conjunction with cyberbullying, and that is the way individuals communicate out of contemplation of group norms.

SIDE theory is a reinvention of deindividuation theory that places more weight on situational circumstances (e.g., Christopherson, 2007). SIDE theory suggests that, when all group members are anonymous, group salience and member identification with the group both intensify (e.g., Spears & Lea, 1992). However, if some group members are identifiable while some are anonymous, SIDE theory suggests that the anonymous members will identify less with the group and more with themselves as individuals. Therefore, anonymous members would be more likely to behave in ways that are harmful to the group (Spears & Lea, 1992). This includes prejudiced rhetoric and aggression, assuming that the group does not directly encourage such conduct. This explains why certain platforms, such as YouTube, where users can choose to be anonymous or

identified, produce more hateful anonymous behavior than websites such as Facebook or Reddit, where virtually all members are identifiable or all members are anonymous. The common association of anonymous commenters with meaningless prejudiced rhetoric and harmful actions might suggest internet users generally attribute anonymous posts to lower levels of honesty and higher levels of attention seeking. However, the type of prejudiced rhetoric or harmful action as well as the individual differences and biases present in third-party observers (e.g., social vigilantism, racial attitudes) may affect the degree to which those observers attribute anonymous posts to honesty and attention-seeking as well as the degree to which those observers condone intervention or intervene themselves.

Social Vigilantism (SV)

Social Vigilantism (SV) is the tendency for individuals to believe their opinions are superior to others' and to attempt to impress their beliefs onto others (Saucier & Webster, 2010). It is important to note that SV is an individual difference that is measured on a continuum. It is a tendency toward believing one's opinions are superior, greater efforts to defend their beliefs against persuasion attempts and challenges, and greater tendencies to attempt to impress those beliefs onto other people. Such impressions of beliefs may take many forms, such as arguing and counterarguing over social media.

SV is associated with more resistance to persuasion when attitudes about sex education (Saucier & Webster, 2010), abortion, the war in Iraq, (Saucier, Webster, Hoffman, & Strain, 2014), and climate change (O'Dea et al., 2018) are challenged. This resistance to persuasion is accompanied by less attitude change following the challenge (Saucier & Webster, 2010); greater use of strategies to resist persuasion such as counterarguing, attempting to impress beliefs onto others and, to a slightly lesser extent, engaging in selective exposure (Saucier & Webster, 2010;

Saucier et al., 2014; O'Dea, Zhu, & Saucier, in preparation; O'Dea et al., 2018); and increased levels of both positive and negative affect in reaction to the challenge (O'Dea et al., 2018). Taken together, this indicates that SV could be associated with engaging in and perhaps enjoying arguing in general, including arguing against inflammatory comments on social media. Additionally, previous research has found that, though SV is not associated with being more well-informed, it is associated with the use of strategies to prevent persuasion, including counterarguing (Saucier, Smith, & Lawless, 2021) and that SV is associated with counterarguing in response to extreme political opinions (e.g., Saucier & Webster, 2010; Raimi & Leary, 2014). SV predicts responses such as counterarguing and resistance to persuasion above and beyond individual differences in dogmatism, narcissism, moral stability, need for cognition, and reactance (Saucier & Webster, 2010), as well as argumentativeness, attitude strength, and the importance of the issue (Saucier et al., 2014). Because SV is related to counterarguing in this way, and because it predicts counterarguing above and beyond other related variables, it makes sense that SV would be associated with perceiving confronting or intervening against online trolls positively and with intervening oneself.

Social Desirability

The validity and worth of psychological assessment relies on accurate responding. Misreporting perceptions, thoughts, and feelings can easily invalidate the results of psychological studies by contributing faulty data. As a result, intentional misreporting represents a significant concern to the field (Ben-Porath & Waller, 1992; Cashel, Rogers, Sewell, & Martin-Cannici, 1995; Nichols & Greene, 1997). This is especially true for studying subjects that are socially taboo or highly controversial.

In the creation of the Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe, 1960), the researchers define social desirability as the need for social approval. Paulhus (1984) suggested that the relevant part of social desirability for most psychological researchers is impression management, which may be undertaken by participants in order to seem "better" to the public or to the researcher. Due to the sensitive nature of prejudice and the general social consequences that usually come with expressing prejudices, some individuals may be motivated to respond in a manner that makes them look non-racist or otherwise "good" as a way to impression manage. For example, it has been found that people may try to present a more favorable impression of themselves, such as by endorsing positive characteristics or behaviors (Bagby et al., 1999). Research has shown that psychological measures can be distorted across an extensive spectrum of settings, from job applications to inpatient units, often while successfully avoiding discovery (e.g. Baer & Miller, 2002; Bagby et al., 1999; Pauls & Crost, 2005; Viswesvaran & Ones, 1999). Social desirability likely motivates the disavowal of negatively perceived personality attributes (Bäckström, Björklund, & Larsson, 2009; Viswesvaran & Ones, 1999). In work settings, participants can successfully simulate socially desirable responses, such as positive attributes for specific job descriptions (Bagby & Marshall, 2003; Furnham, 1990; Pauls & Crost, 2004; Retzlaff, Sheehan, & Fiel, 1991; Scandell & Wlazelek, 1996). Social desirability may especially be an issue in discussions of racism.

Individuals may be unwilling to reveal undesirable personality traits, such as racism, because of the social pressures associated with such traits (Martin, Pescosolido, & Tuch, 2000). Many White people, for example, may be concerned about appearing prejudiced in interactions with or regarding Black people (e.g., Butz & Plant, 2006; Plant, 2004; Shelton, 2003). Concerns about the transparency of prejudice may heighten anxiety and interfere with people's ability to

convey unbiased impressions in interracial interactions (e.g., Plant & Butz, 2006). In discussing expressions of and responses to prejudice, it is critical to consider not only whether people are motivated to respond without prejudice but also the reasons underlying their motivation (Dunton & Fazio, 1997; Plant & Devine, 1998). The Justification-Suppression Model (JSM; Crandall & Eshleman, 2003) of intergroup bias provides a valuable framework for understanding the motivation to avoid expressing prejudice. According to the JSM, although intergroup bias can be freely expressed in some contexts, intergroup biases are often socially undesirable, and therefore expressions of intergroup bias tend to be suppressed. Within contexts in which an intergroup bias is generally suppressed, biases are subsequently expressed to the extent that one can justify (or legitimize) the bias. Expressions of prejudice (and, perhaps by extension, failures to fight prejudice) may come with social sanctions that may motivate individuals to respond without prejudice. These social sanctions may be especially salient in the current online environment due to the phenomenon of cancel culture (e.g., Ng, 2020; Romano, 2019). This indicates that, especially when discussing socially fraught topics (e.g., racial prejudice) or when individuals perceive that there may be negative social consequences to expressions of their true feelings, people may self-police their responses.

Additionally, social desirability is inversely related to the degree of privacy and anonymity that a person experiences (Ben-Ze'ev, 2003; Buchanan, 2000; Davis, 1999; Fisher, 1993; Joinson, 1999; Pasveer & Ellard, 1998; Smith & Leigh, 1997). It has been argued that computers offer limited social context, making the user feel anonymous and self-absorbed (Sproull & Kiesler, 1986). Because the current studies discuss socially sensitive topics in an online environment, it is difficult to predict the effect that social desirability may have on responses. To address the vulnerability of the current studies to response distortion, the

Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe, 1960) will be administered to detect social desirability. This scale is based on detection strategies, for example, the assumption that participants who score significantly above the norm on items about socially desirable qualities might be exaggerating their positive self-presentation.

Need for Chaos

Need for Chaos is defined as a desire for the destruction of order and established structures (Arceneaux, Gravelle, Osmundsen, Petersen, Reifler, & Scotto, 2021; Petersen, Osmundsen, & Arceneaux, 2018). This destruction may take the form of many different destructive behaviors, including support for and instigation of large-scale intergroup conflict (e.g., along racial or gender-based lines). Although multiple psychological motivations shape the spread of prejudice and stereotypes in general (DiFonzo & Bordia, 2007), some evidence suggests that the sharing of negative stereotypes about other groups specifically relates to states of conflict between the target group and the group of the rumor sharer (Laustsen & Petersen, 2015). In this viewpoint, the person posting the antagonistic comments is less concerned with the truth and more concerned with the value of the rumor to aiding in their side "winning" the conflict. Additionally, people who post aggressive rumors and stereotypes may be motivated by what Petersen, Osmundsen, and Arceneaux (2018) call "chaotic" motivations. That is, when people share antagonistic rumors and stereotypes, they might do so in order to mobilize spectators against the entire social order, rather than aiding one particular group against another. It is possible that someone with such a need for chaos may have sympathy for "trolling" behaviors such as instigating prejudicial arguments, and therefore may view "trolling" behavior as acceptable or as less morally reprehensible. Therefore, the Need for Chaos Scale (Petersen, et al., 2018) will be implemented to measure need for chaos and control for this possible sympathy.

Racial Attitudes

The fundamental nature of White North American attitudes towards Black people as overtly negative is largely considered no longer be socially acceptable (e.g., Linder, 2015; Sue, 2013). Unfortunately, negative attitudes based on race have not been eliminated, but have only grown more nuanced and well-hidden. Obvious discriminatory behaviors and prejudices are frowned upon, and people are anxious to avoid behaving in a manner that could be interpreted as biased or prejudiced (Fiske, 1998; Gaertner & Dovidio, 1986; Plant & Devine, 1998). One's personal prejudices or biases, however, can be expressed in far more subtle ways (e.g., Gaertner & Dovidio, 1981; Miller, O'Dea, Lawless, & Saucier, 2019; Saucier & Miller, 2003). Although most individuals in contemporary North American society face a strong societal and cultural demand to endorse egalitarian principles, discrimination still exists (e.g., Saucier, Miller, Martens, & O'Dea, 2017). It has been demonstrated in laboratory settings, such as in the case of helping behavior in both emergency (e.g., Gaertner & Dovidio, 1977) and nonemergency situations (e.g., Gaertner & Dovidio, 1986; McManus, Saucier, O'Dea, & Bernard, 2019; Saucier, Miller, & Doucet, 2005) as well as in attitudes regarding crime and police violence (e.g., Miller, O'Dea, & Saucier, 2021; Miller, Peacock, & Saucier, 2021; Saucier, Hockett, O'Dea, & Miller, 2016; Saucier, Hockett, & Wallenberg, 2008; Saucier, Hockett, Zanotti, & Heffel, 2010). Although behavior that is overtly prejudiced or discriminatory is socially unacceptable and most individuals therefore consciously avoid and control explicit expression of prejudice in their responses in interracial situations, implicit and more subtle biases are still common (e.g., Devine, 1989, Fairchild, 2018; Greenwald, McGhee & Schwartz, 1998; Johnson-Agbakwu, Ali, Oxford, Wingo, Manin, & Coonrod, 2020; Payne & Hannay, 2021).

Implicit prejudice has been shown to influence the inequitable behavior of aversive racists (Son Hing, Chung-Yan, Grunfeld, Robichaud & Zanna, 2005), and also to predict the level of prejudice that independent observers and Black confederates themselves perceive in the nonverbal behaviors of White participants during interracial interactions in the lab (Dovidio, Kawakami, & Gaertner, 2002). Large inconsistencies between White individuals' positive explicit attitudes and negative implicit attitudes toward Black people are therefore common. To attempt to control for this potential bias, two measures of racial attitudes will be employed.

Miller and Saucier (2018) created the Propensity to Make Attributions to Prejudice Scale (PMAPS) to assess the tendency to attribute behaviors to bias. The PMAPS has been shown to predict attributions to prejudice in a variety of situations, particularly unclear situations where behavior may be attributed to factors unrelated to race (e.g., Miller et al., 2021; Miller et al., 2017; Sparrow, 2019; Stratmoen, Lawless, & Saucier, 2019) or when the prejudice is obscured by other factors, such as humor (e.g., Miller et al., 2019). Additionally, PMAPS may be negatively associated with motivations to keep the existing social hierarchy and with anger when historically lower-status groups (i.e., Black people) claim discrimination (Miller et al., 2017). The ambiguity of online and/or anonymous situations as well as the potential threat the internet poses to the social hierarchy (e.g., Keum & Miller, 2018) makes PMAPS an apt scale to measure the tendencies of participants to attribute racially prejudiced posts to prejudice within the person posting them as opposed to other potential factors like attention seeking. The Attitudes Toward Blacks (ATB) Scale was constructed by Brigham (1993) to measure racial attitudes toward Black people in four central areas, including feelings of social distance or disquiet when interacting with Black people, negative emotional reactions to Black people, administrative or institutional policy (e.g., open housing, equality), and personal apprehension about being denied a job or

promotion due to special treatment for Black people (based on affirmative action programs). The ATB Scale has been used recently in studies as a measure of prejudice in social mobility, health, and economic policies (e.g., Agadjanian, Carey, & Horiuchi, 2021; Bianchi, Hall, & Lee, 2018; DeSante & Smith, 2020; Jardina & Piston; 2019; Mandalaywala, Amodio, & Rhodes, 2018; Milner & Franz, 2020; Yadon & Piston, 2019) and in studies that demonstrate the effectiveness of third-party confrontation on reducing prejudice (e.g., Czopp, Monteith, & Mark, 2006). The ATB does come with the limitation that it is originally meant to measure White people's responses to Black people, and the participants in the current studies were not all White (though a majority were). Because the majority of participants were White, and because the ATB is much more overt than the PMAPS as well as the recent use of the ATB in studies on the confrontation of prejudice, it is included in the following studies, which examine perceptions of overtly racist posts and the comments confronting them. Both the PMAPS and the ATB were used to cover multiple types of expressions of prejudice and to cover the diverse racial makeup of the participant pool.

The Lawless Model of Perceiving Trolling

My past research findings on trolling in social media can be expressed by the following Lawless Model of Perceiving Trolling (see Figure 1). Essentially, the perceived prejudice of a post, the anonymity of the Original Person (OP) who posted it, and the target of the post being a group of people influence third-party perceptions of trolling (i.e., decreased perceptions that the post reflects what the OP really believes, increased perceptions of the OP as attention-seeking, and higher ratings of the OP as a "troll"). Additionally, the third-party's individual PMAPS score has also been shown to be positively associated with perceiving posts as prejudiced. The current research will further test this model as well as extend it.

The current research adds a new outcome variable, confrontation of the post by the third party, to the model (see Figure 2). It is hypothesized that perceived trolling will generally be negatively associated with confrontation of the post. This would make sense because perceived trolling indicates that the third party does not think the OP truly believes the prejudiced ideas contained in the post. If the OP does not really believe what they are saying, there may be less reason to confront them. However, it is hypothesized that PMAPS scores will be positively associated with confrontation of the post regardless of perceived trolling. PMAPS is associated with perceiving prejudice more readily and with confronting it more often. Therefore, it is likely that PMAPS will be positively associated with confronting prejudiced posts regardless of the perceived intention of the OP. Additionally, the current research adds a new individual difference predictor to the model, Social Vigilantism (SV). Like PMAPS, it is also hypothesized that SV will be positively associated with confrontation of prejudiced posts regardless of perceptions of trolling. SV indicates a desire to impress one's own beliefs onto others and is associated with counterarguing and other strategies that indicate a willingness to confront statements one disagrees with. Given that prejudice is generally not socially acceptable, it is reasonable to expect that most participants will disagree with the content of extremely prejudiced posts. Therefore, SV should be associated with confronting those posts.

Overview of Current Studies

While past studies have focused on how anonymity of both the perpetrator and target affects third-party perceptions of racial prejudice online (Lawless & Saucier, submitted a; submitted b; in preparation), little is known about how individual differences within the third-party affect perceptions of such online interactions. It is possible that individual differences, including social vigilantism and personal endorsement of prejudicial beliefs, affect how third

parties perceive and would choose to interact with prejudicial statements online. Social vigilantism (SV) is the tendency for individuals to believe their opinions are superior to others', and to attempt to impress their beliefs onto others (Saucier & Webster, 2010). This impression of beliefs may take many forms, such as arguing over social media. Previous research has found that, though SV is not associated with being well-informed, it is associated with the use of strategies to prevent persuasion, including counterarguing (Saucier et al., 2021) and that SV is associated with counterarguing in response to extreme political opinions (e.g., Saucier & Webster, 2010; Raimi & Leary, 2014). SV predicts responses such as counterarguing and resistance to persuasion above and beyond individual differences in dogmatism, narcissism, moral stability, need for cognition, and reactance (Saucier & Webster, 2010), as well as argumentativeness, attitude strength, and the importance of the issue (Saucier et al., 2014). Because SV is related to counterarguing in this way, it makes sense that SV would be associated with viewing confronting prejudice online as a positive behavior.

Another individual difference that may affect views of arguing against prejudice online are participants' own prejudicial attitudes. Miller and Saucier (2018) created the Propensity to Make Attributions to Prejudice Scale (PMAPS) to assess the tendency to attribute actions to prejudice. The PMAPS has been shown to predict attributions to prejudice in a variety of situations (e.g., Miller et al., 2017; Stratmoen et al., 2019). Because extremely prejudiced statements online can be attributed to either true prejudice or to trolling behaviors, it makes sense that PMAPS may affect views of such statements. The Attitudes Toward Blacks (ATB) Scale was constructed by Brigham (1993) to measure White people's racial attitudes toward Black people and has been used as a measure of prejudice in studies that demonstrate the effectiveness of third-party confrontation on decreasing prejudice (e.g., Czopp, Monteith, & Mark, 2006). The

proposed studies explore how these individual differences and anonymity affect third-party views of confrontations to prejudice. The potential implications of the current studies may be that factors of the online environment, such as anonymity of platform, as well as individual differences affect how individuals react to prejudicial speech online. These and future studies along this line of research are important to fully understand the factors at play within internet culture.

Chapter 2 - Study 1

Study 1 examines the attributions and judgements people make in response to confrontations of online prejudice in both anonymous and identifiable conditions. Several displays of aggressive and overt racism were created and placed those displays in a particular social media environment (i.e., anonymous or identifiable) similar to Lawless and Saucier (submitted a; submitted b; in preparation). Additionally, these posts were either displayed with no other interactions or with a single other commenter who confronts and argues back and forth with the Original Person who posted (the OP). Participants saw these posts in a counterbalanced within-subjects 2(anonymity) x 2(including commenter arguments) design. Participants also filled out individual difference measures of SV, PMAPS, and ATB.

In Study 1, items for participants to rate how racist they find the post, how racist they find the OP, to what extent they believe OP agrees with the sentiments expressed, and to what extent they believe the post was made to garner attention or disturb the status quo were also created. These items were based on Lawless and Saucier (submitted a; submitted b). While the effects of anonymity on these perceptions have been documented, it is likely that including a back-and-forth confrontation or argument may change these perceptions. It is possible that such an argument will make participants see the OP as more honest, or perhaps as more of an attention-seeker who simply likes to argue (i.e., a troll). Additionally, PMAPS and ATB are expected to be associated with these perceptions such that participants who display more prejudicial beliefs may rate prejudicial statements as less racist and more honest. In addition to these items, new items that measure the extent to which participants feel the confronting commenter did the right thing by confronting were also included. It is possible that a confronting commenter could be seen as standing up rightly against prejudice, but it is also possible that such a commenter could be seen as "feeding the trolls"

and giving the OP what they wanted, thus reinforcing the prejudicial behavior. Because SV is associated with counterarguing, it is possible that SV is associated with seeing confronting as the right thing to do.

Method

Participants

Participants consisted of 234 volunteers who were recruited from Amazon's Mechanical Turk software and participated in exchange for a small amount of monetary compensation (i.e., $0.25). However, of these 234 participants, 28 were removed for finishing the survey in less than three minutes and 14 were removed for failing the attention checks and/or bot captcha; therefore, 192 participants' responses were analyzed. An a priori power analysis (gPower) with an $\alpha = .05$ and power of .95 was conducted. Further, the effect size which was entered into gPower was taken from Lawless and Saucier (submitted b), which showed effect sizes of approximately .20. This analysis yielded an approximate sample size of 117 participants necessary to achieve the boundaries discussed. The participants were 63.5% men, and 33.3% women, with 6 participants identifying as outside the gender binary. Participants were 71.3% White, 13.5% Black, 4.7% Latino/a, 6.2% Asian, 0.5% Native American, and 3.6% Multiracial. Participants' ages ranged from 18 to 76, with a mean of 37.71. To ensure participant anonymity, participant names were not collected and worker identification numbers were kept separately from all other study materials. Identification information was only collected for the purposes of informed consent and exchanging appropriate compensation.

Mock Social Media Posts.

Stimuli similar to Lawless & Saucier (submitted b) were used. In their studies, Lawless and Saucier used overtly racist statements targeting groups of people and manipulated the

anonymity of the person posting the statements by placing the statements in mock Facebook (identifiable) and Reddit (anonymous) posts. Reddit is a website where users post content behind any screenname they want, and it does not require any link to an identifiable account of any kind. Thus, though there is a screenname that accompanies posts on Reddit, it is virtually impossible to connect a non-identifiable screenname to a real-life person, and posts are therefore almost entirely anonymous.

The current study examined whether perceptions of racist posts differ based on both anonymity and the presence or absence of dissenting comments. Therefore, 8 overtly racist mock social media posts that attack a group of people similar to those used by Lawless and Saucier (submitted b; e.g., *Black people whine and complain about being "oppressed" yet sit at home and collect welfare. It's called hard work*) were specifically used, and those posts appeared either without any feedback at all or including confronting comments (e.g., *You should be banned for saying stuff like this*; see Appendix B for additional examples). Posts were evenly split amongst identifiable (Facebook) and anonymous (Reddit) social media platforms and were counterbalanced such that all participants saw 2 posts from both interaction conditions in each anonymity condition in a within-subjects design.

Individual Differences.

Social Vigilantism. To measure the extent to which individuals believe their beliefs are superior to others' and seek to impose those beliefs onto others, the Social Vigilantism (SV) Scale (Saucier & Webster, 2010) was used. This scale includes 14 items measured on a 1 (*strongly disagree*) to 9 (*strongly agree*) Likert-type scale. It includes items such as *I feel as if it is my duty to enlighten other people*. A composite score for SV was calculated by reverse-

scoring antithetical items and calculating an average score for each participant with higher scores indicating greater tendencies to impress one's beliefs onto others.

Propensity to Make Attributions to Prejudice. To measure beliefs about the prevalence of racial prejudice, the Propensity to Make Attributions to Prejudice Scale (PMAPS; Miller & Saucier, 2016) was used. The scale includes 15 items measured on a 1 (*strongly disagree*) to 9 (*strongly agree*) Likert-type scale. It includes items such as *I consider whether people's actions are prejudiced or discriminatory.* A composite score for PMAPS was calculated by reverse-scoring antithetical items and calculating an average score for each participant with higher scores indicating greater tendencies to attribute causes of behavior to racial prejudice.

Explicit prejudice toward Black People. To measure participants' level of explicit racial prejudice toward Black individuals, the Attitudes Toward Blacks (ATB; Brigham, 1993) scale was used. The scale includes 20 items measured on a 1 (*strongly disagree*) to 9 (*strongly agree*) scale. It includes items such as *I would rather not have Blacks live in the same apartment building I live in.* A composite score for ATB was calculated by reverse-scoring antithetical items and calculating an average score for each participant with higher scores indicating greater levels of blatant anti-Black prejudice.

Social Desirability. To measure participants' tendencies toward socially desirable behavior, the Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe, 1960) was used, which defines social desirability as the need for social approval. This instrument includes 33 items, which are to be classified as true or false by the respondent. Some of these items correspond to sentences that describe desirable but uncommon daily behaviors (attribution items, scored if answered "true"; e.g., *I am always courteous, even to people who are disagreeable.*), whereas others describe highly common but socially undesirable behaviors (denial items, scored

when answered "false"; e.g., *There have been occasions when I felt like smashing things*). Therefore, social desirability is scored from 0 – 33 as the number of socially desirable responses made by the participant.

Need for Chaos. To measure participants' levels of desire to fight against established social order, the Need for Chaos Scale (Petersen et al., 2018) was used. The scale includes eight items measured on 1 (*strongly disagree*) to 9 (*strongly agree*) scale. It includes items such as *I think society should be burned to the ground.* A composite score for Need for Chaos was calculated by calculating an average score for each participant with higher scores indicating greater levels of desire for chaos.

Criterion variables

Each of the following measures characterizes a specific facet of perceptions of online prejudice and the people posting it that has been discussed in previous literature. Specifically, items assessing the extent to which participants perceive each post as racist and honest as well as items assessing the extent to which participants perceive the OP as racist, honest, genuine in their belief of what they have posted, attempting to convince others of what they have posted, and attempting to seek attention for attention's own sake were included. Additionally, on posts that contain confronting comments, items assessing the perception that the commenter did the right thing by confronting were included. Each of these measures is described below, and the materials are included in Appendix C.

Perceived racial prejudice of the post. To examine the extent to which participants perceive each post as racist, a perceived racial prejudice item employed by Lawless and Saucier (submitted b) was used. This item is measured on a 1 (*not at all*) to 9 (*very much*) scale, with higher ratings indicating greater levels of perceived racial prejudice of the post.

Perceived maliciousness of the post. To examine the extent to which participants perceive each post as malicious, 3 perceived maliciousness items (e.g., *This post is meant to harm*) used in previous research (e.g., Lawless & Saucier, b) were used. Each item is measured on a 1 (*not at all*) to 9 (*very much*) scale. A composite score for the perceived maliciousness of the post was calculated by reverse scoring antithetical items and calculating an average score with higher scores indicating greater levels of perceived maliciousness of the post.

Perceived honesty of the post. To examine the extent to which participants perceive each post as honest, 2 perceived honesty items similar to those used by Lawless and Saucier (submitted b) were used. Each item is measured on a 1 (*not at all*) to 9 (*very much*) scale. A composite score for the perceived honesty of the post was calculated by calculating an average score with higher scores indicating greater levels of perceived honesty of the post.

Perceived racial prejudice of the OP. To examine the extent to which participants perceive the OP as racist, a perceived racial prejudice item employed by Lawless and Saucier (submitted b) was used. This item is measured on a 1 (*not at all*) to 9 (*very much*) scale, with higher ratings indicating greater levels of perceived racial prejudice of the person posting the statement.

Perceived maliciousness of the OP. To examine the extent to which participants perceive each OP as malicious, 3 perceived maliciousness items (e.g., *The person who posted this intended to harm the person(people) this post is about.*) were used. Each item is measured on a 1 (*not at all*) to 9 (*very much*) scale. A composite score for the perceived maliciousness of the OP was calculated by reverse scoring antithetical items and calculating an average score with higher scores indicating greater levels of perceived maliciousness of the OP.

Perceived honesty of the OP. To examine the extent to which participants perceive each OP as honest, 2 perceived honesty items similar to those used by Lawless and Saucier (submitted b) were used. Each item is measured on a 1 (*not at all*) to 9 (*very much*) scale. A composite score for the perceived honesty of the OP was calculated by calculating an average score with higher scores indicating greater levels of perceived honesty of the OP.

Perceived attention seeking of the OP. To examine the extent to which participants perceive each OP as seeking attention *rather than* being honest, a perceived attention-seeking item similar to that employed by Lawless and Saucier (submitted b; i.e, *The person who posted this is simply looking for attention*) was used. This item is measured on a 1 (*not at all*) to 9 (*very much*) scale, with higher ratings indicating greater levels of perceived attention-seeking of the OP.

Perceived righteousness of the commenter. To examine the extent to which participants perceive each confronting commenter as doing the right thing, 2 new items (e.g., *The commenter who confronted this post did the right thing*) were used. Each item is measured on a 1 (*not at all*) to 9 (*very much*) scale. A composite score for the perceived righteousness of the commenter was calculated by calculating an average score with higher scores indicating greater levels of perceived righteousness.

Procedure

The current study was conducted online using Amazon's Mechanical Turk software. Once participants signed up, they followed a link to the study on Qualtrics. Participants gave informed consent prior to participation. After providing demographic information (e.g., sex, race, age; see Appendix A for materials), participants read and responded to all 12 mock social media posts in the counterbalanced fashion described above. Participants were debriefed after they

completed the study to allow the experimenters to answer any questions the participants may have had.

Results

Following the cleaning of the dataset (e.g., removing participants who completed the questionnaire in an unrealistic amount of time, removing participants who failed the bot captcha), composite scores for each of the continuous variables were computed. As noted in the Materials section, for each of the measures, participants' scores on each individual item were averaged after reverse scoring antithetical items to create composite scores. On each measure, higher scores represent higher levels of the construct being measured.

The bivariate correlations among the predictor variables (see Table 1) were examined. Consistent with previous research, PMAPS and ATB were significantly negatively correlated and PMAPS and SD were not significantly correlated (e.g., Miller & Saucier, 2018). Additionally, social desirability was negatively correlated with ATB, which was expected because it is not widely socially acceptable to be blatantly racist. However, these correlations are not central to the main hypotheses of the current studies, so they will not be discussed further.

The bivariate correlations among the predictor variables (see Table 2) and the participant demographics of gender and ethnicity were also examined. Because of a lack of diversity within the sample, ethnicity was collapsed into two categories: White and of Color. Similarly, it was necessary to collapse gender into only the categories man and woman while removing participants who identified outside of that binary from this portion of the analysis. Consistent with previous research, PMAPS and ATB were significantly correlated with ethnicity such that participants of Color were less racist themselves and more likely to make attributions to prejudice (e.g., Miller & Saucier, 2018). This makes sense because many of the participants of

Color were Black, and the ATB specifically measures racism against Black people. There were no significant correlations with participant gender in this study. However, these correlations are not central to the main hypotheses of the current studies, so they will not be discussed further.

The bivariate correlations among the criterion variables were then examined: perceived racial prejudice of the post, maliciousness of the post, honesty of the post, racial prejudice of the OP, maliciousness of the OP, honesty of the OP, attention seeking of the OP, and righteousness of the commenter (see Table 3). Consistent with previous research by Lawless and Saucier (submitted b), there were positive correlations between the perceived racial prejudice of the post, maliciousness of the post, racial prejudice of the OP, and maliciousness of the OP. There was also a positive correlation between perceived honesty of the post and honesty of the OP. In addition, there was a negative correlation between perceived honesty of the OP and perceived attention seeking of the OP, which is consistent with past research and with the Trolling Hypothesis. Furthermore, perceived righteousness of the confronting commenter was also positively correlated with perceived honesty of the post and OP, which is consistent with the Trolling Hypothesis in that it may indicate that arguing with posts perceived as trolling is seen as the wrong thing to do, as "feeding the trolls" and reinforcing the trolling behavior.

Whether anonymous posts are seen as less racist, less honest, and more attention seeking than similar identifiable posts as well as the effects of the presence of confrontational comments on the posts were then tested. Recall, there are two competing hypotheses founded on previous research. The first hypothesis, the Trolling Hypothesis, states that anonymous statements will be rated as less honest and prejudiced, and more attention-seeking than identifiable statements because the statements are thought of as being provoked not by genuine feeling, but by the thrill of being allowed to broadcast socially unacceptable statements without the consequences that

would come with being identifiable. This would also lead to engaging with trolling posts via confrontational comments being seen as less appropriate because it is seen as giving the troll what they want. The second hypothesis, the Disinhibition Hypothesis, states that anonymous statements will be rated as more honest and prejudiced, and less attention-seeking than identifiable statements because anonymity is thought of as removing the social pressures that usually inhibit genuine hate speech. This would also indicate that confrontational comments would be perceived positively because they are perceived as arguing against honest prejudice.

To test these hypotheses against one another, a series of repeated measures multilevel model analyses was conducted predicting the criterion variables and including SV, PMAPS, NFC, Anonymity, Inclusion of Comments, their two-way interactions, and the three-way interaction between Anonymity, SV, and Inclusion of Comments as predictor fixed effects, and allowing participants' intercepts to vary (see Tables 4-12). Consistent with the hypotheses, there were significant unique effects of PMAPS (*Prejudice of Post: F*(1, 186) = 6.06, $p < .001$; *Prejudice of Person: F*(1, 186) = 8.70, $p < .001$, *Righteousness of Commenter F*(1, 186) = 2.99, $p = .002$; see PMAPS β values in Tables 4-11) such that, generally people higher in PMAPS viewed posts and people as both more prejudiced and more malicious and viewed commenters as doing the right thing. Also, there were significant unique effects of Need for Chaos (*Maliciousness of Post*: *F*(1, 186) = 18.07, $p < .001$, *Maliciousness of Person F*(1, 186) = 3.62, $p < .001$; see Need for Chaos β values in Tables 4-11) such that, generally people higher in Need for Chaos viewed posts and people as less malicious, perhaps because need for chaos is associated with wanting to buck the social order, potentially leading to sympathizing with online behavior that does so. There were additional significant unique effects of SV (*Righteousness of Commenter*: *F*(1, 186) = 4.63, $p < .001$; see SV β values in Tables 4-11) such that people higher

in SV rated the confronting commenter as generally doing more of the right thing. This is consistent with the associations SV has with counterarguing, and demonstrates that SV may be associated with perceiving counterarguing when one disagrees as the right thing to do.

Additionally, consistent with the Trolling hypothesis, there were significant unique effects of Anonymity (*Prejudice of Person: F*(1, 445) = 6.24, *p* = .004, *Maliciousness of Person F*(1, 445) = 5.64, *p* = .001, *Honesty of Person: F*(1, 445) = 4.36, *p* = .018, *Honesty of Post: F*(1, 445) = 3.57, *p* = .002; *Righteousness of Commenter: F*(1, 445) = 2.54, *p* = .018; see Anonymity β values in Tables 4-11), such that people posting anonymous posts were rated as less prejudiced, malicious, and honest than people posting identifiably, and such that commenters on anonymous posts were rated as doing less of the right thing. This could suggest that people view anonymous posts as trolling, not meant to be taken seriously or as truth, but rather intended to garner extreme reactions by bucking against the social order. There were also significant unique effects of presence of Comments (*Prejudice of Post: F*(1, 445) = 3.46, *p* < .001; *Maliciousness of Post*: *F*(1, 445) = 4.52, *p* < .001, *Prejudice of Person: F*(1, 445) = 3.53, *p* < .001, *Maliciousness of Person F*(1, 445) = 4.52, *p* < .001; see Comments β values in Tables 4-11) such that posts with confrontational comments were rated as generally more malicious and more prejudiced. This may be because the presence of a confronting comment indicates that another person believed the post was malicious and prejudiced enough to warrant an argument. These main effects were qualified by significant two-way interactions between Anonymity and presence of Comments (*Honesty of Person: F*(1, 445) = 4.54, *p* = .006, *Honesty of Post: F*(1, 445) = 2.97, *p* = .003, *Attention Seeking*: *F*(1, 445) = 2.20, *p* = .029, see interaction term β values in Tables 4-11). These interactions indicate that the effects of anonymity of post on these criterion variables depended upon whether the post included confrontational arguments from others.

Simple slopes analyses were then conducted on the interaction terms that were significant to determine whether the final hypotheses were supported (see Table 12). As predicted, identifiable posts that included confrontational comments were rated as more honest and less attention-seeking than anonymous posts regardless of the presence of commenters. This is consistent with the Trolling Hypothesis, which predicted that anonymous posts would be seen as less honest and more attention seeking and that comments on such anonymous posts would be seen as evidence of "feeding the trolls."

Additionally, to test the viability of the Lawless Model of Perceiving Trolling, a path analysis was conducted on the parts of the model relevant to Study 1 (see Figure 3). Consistent with past research (Lawless & Saucier a) and with the hypotheses, the perceived prejudice of a post and the anonymity of the Original Person (OP) who posted it influence third-party perceptions of trolling (i.e., decreased perceptions that the post reflects what the OP really believes, increased perceptions of the OP as attention-seeking, and higher ratings of the OP as a "troll"). Additionally, the third-party's individual PMAPS score was shown to be positively associated with perceiving posts as prejudiced. Though SV was included as a predictor in Study 1, it did not significantly uniquely predict any outcome variables except for Righteousness of the Commenter, and therefore its place in this model is currently unclear. However, Study 2 adds a new outcome variable, Confrontation of the Post, to the model, which should help clarify the importance of SV as an individual difference.

Chapter 3 - Study 2

Study 2 used the same methods as Study 1, but this time participants were asked how they would interact with the posts and gave them a chance to write their own comment on the posts. Therefore, participants in Study 2 were evaluating interactions as potential interactors rather than as third-parties. Similar results as Study 1 are expected to the items measuring perceptions of the OP since those perceptions may not be affected by the opportunity to participate. PMAPS and ATB are expected to be associated with interactions such that anti-prejudice attitudes will be associated with more confrontation of prejudice via interactions. SV is also expected to be associated with more confrontation of prejudice and leaving more confronting comments because SV is associated with the tendency to counterargue. It is also possible that anonymity will affect the results such that posts that are identifiable may receive more interactions because they are generally seen as more honest, which would be consistent with the Trolling Hypothesis, or such that anonymous posts may receive more interactions because participants may be emboldened by their own anonymity, which would be more consistent with the Disinhibition Hypothesis. Similarly, posts that already have confronting comments on them may be interacted with less because someone else has already confronted the OP or may be interacted with more because participants may feel more comfortable taking an action someone else already clearly agrees with.

Method

Participants

Participants consisted of 255 volunteers who were recruited from Amazon's Mechanical Turk software and participated in exchange for a small amount of monetary compensation (i.e., $0.25). However, of these 256 participants, 47 were removed for finishing the survey in less than three minutes and 15 were removed for failing the attention checks and/or bot captcha; therefore,

193 participants' responses were analyzed. An a priori power analysis (gPower) was conducted with an $\alpha = .05$ and power of .95. Further, the effect size which was entered into gPower was taken from Lawless and Saucier (submitted b), which showed effect sizes of approximately .20. This analysis yielded an approximate sample size of 117 participants necessary to achieve the boundaries discussed. The participants were 64.7% men, and 33.2% women, with 4 participants identifying as outside the gender binary. Participants were 69.9% White, 13.0% Black, 6.7% Latino/a, 7.7% Asian, and 2.6% Multiracial. Participants' ages ranged from 19 to 75, with a mean of 38.85. To ensure participant anonymity, participant names were not collected, and worker identification numbers were kept separately from all other study materials. Identification information was only collected for the purposes of informed consent and exchanging appropriate compensation.

Mock Social Media Posts. The content from the same 12 overtly racist mock social media posts that were used in Study 1 were again used in Study 2. Content was presented in a 2 (identifiable/anonymous platform) x 2 (not/containing confronting comments) within-subjects design. That is, posts were evenly split amongst identifiable (Facebook) and anonymous (Reddit) social media platforms and were counterbalanced such that all participants saw posts from both interaction conditions in each anonymity condition. Posts and messages appeared to be on either an identifiable (Facebook) or anonymous (Reddit) social media platform and appeared either free of interactions or including confronting comments. All manipulations were presented to participants in a counterbalanced fashion.

Criterion variables. Criterion variables were the same as those in Study 1. These variables included: perceived racial prejudice of the post, perceived maliciousness of the post, perceived honesty of the post, perceived racial prejudice of the OP, perceived maliciousness of

the OP, perceived honesty of the OP, perceived attention seeking of the OP, and perceived righteousness of the confrontational commenter. Each of these measures was chosen to represent a specific facet of perceptions of online prejudicial speech and the people posting it that has been discussed in previous literature. Materials are included in Appendix C.

Additionally, to examine the extent to which participants would interact with each post, six new items were used (i.e., *I would "like"/upvote this post., I would angry face react/downvote this post., I would leave a positive comment on this post., I would leave a negative comment on this post., I would block the user who made the original post., I would report the user who made the original post.*). Each item is measured on a 1 (*not at all*) to 9 (*very much*) scale. Each of these items was treated as its own separate criterion variable. A text box for participants to type any comments they would like to leave on the post was also included.

Procedure

The current study was conducted online using Amazon's Mechanical Turk software. Once participants signed up, they followed a link to the study on Qualtrics. Participants gave informed consent prior to participation. After providing demographic information (e.g., sex, race, age), participants read and responded to all 12 mock social media posts in the counterbalanced fashion described above. Participants were debriefed after they completed the study to allow the experimenters to answer any questions the participants had.

Results

Following the cleaning of the dataset (e.g., removing participants who completed the questionnaire in an unrealistic amount of time, removing participants who failed the bot captcha), composite scores for each of the continuous variables were computed. As noted in the Materials section, for each of the measures, participants' scores on each individual item were

averaged after reverse scoring antithetical items to create composite scores. On each measure, higher scores represent higher levels of the construct being measured.

I examined the bivariate correlations among the predictor variables (see Table 13). Consistent with previous research and with Study 1, PMAPS and ATB were significantly negatively correlated and PMAPS and SD were not significantly correlated (e.g., Miller & Saucier, 2018) while social desirability was negatively correlated with ATB. However, these correlations are not central to the main hypotheses of the current studies, so they will not be discussed further.

As in Study 1, the bivariate correlations among the predictor variables (see Table 14) and the participant demographics of gender and ethnicity were also examined. Again, because of a lack of diversity within the sample, ethnicity was collapsed into two categories: White and of Color. Similarly, it was also again necessary to collapse gender into only the categories man and woman while removing participants who identified outside of that binary from this portion of the analysis. Consistent with previous research and with Study 1, PMAPS and ATB were significantly correlated with ethnicity such that participants of Color were less racist themselves and more likely to make attributions to prejudice (e.g., Miller & Saucier, 2018). Interestingly, in this study, there were small but significant correlations with participant gender and PMAPS such that women were slightly more likely to make attributions to prejudice than men. This may be due to the fact that women are also more likely to experience discrimination (though of a different kind) than are men, and therefore may be more sensitive to discrimination as a cause of events. However, these correlations are not central to the main hypotheses of the current studies, so they will not be discussed further.

I then examined the bivariate correlations among the criterion variables: perceived racial prejudice of the post, maliciousness of the post, honesty of the post, racial prejudice of the OP, maliciousness of the OP, honesty of the OP, attention seeking of the OP, righteousness of the commenter, self-reported likelihood of leaving a positive interaction, and self-reported likelihood of leaving a negative interaction (see Table 15). Consistent with previous research by Lawless and Saucier (submitted b) as well as with Study 1, there were positive correlations between the perceived racial prejudice of the post, maliciousness of the post, racial prejudice of the OP, and maliciousness of the OP. There was also a positive correlation between perceived honesty of the post and honesty of the OP. In addition, there was a negative correlation between perceived honesty of the OP and perceived attention seeking of the OP, which is consistent with past research and with the Trolling Hypothesis. Furthermore, perceived righteousness of the confronting commenter and likelihood to negatively interact were also positively correlated with perceived honesty of the post and OP, which is consistent with the Trolling Hypothesis in that it may indicate that arguing with posts perceived as trolling is seen as the wrong thing to do, as "feeding the trolls" and reinforcing the trolling behavior.

I then tested whether anonymous posts are seen as less racist, less honest, and more attention seeking than similar identifiable posts, which would be consistent with the Trolling Hypothesis, as well as the effects of the presence of confrontational comments on the posts. Again, it was predicted that anonymous posts would be rated as less honest and prejudiced and more attention-seeking than identifiable comments because it is seen as the person posting not as an act of genuine hatred, but because it is thrilling to participate in the taboo act of espousing prejudicial speech. This would also lead to engaging with trolling posts via confrontational comments being seen as less appropriate because it is seen as giving the troll what they want.

However, it is possible that negative interactions that do not give the troll attention (i.e., blocking or reporting the user) may still be seen a favorable regardless of anonymity or other possible indicators of trolling.

To test these hypotheses against one another, a series of repeated measures multilevel model analyses was conducted predicting the criterion variables and including SV, PMAPS, Anonymity, Inclusion of Comments, their two-way interactions, and the three-way interaction between Anonymity, SV, and Inclusion of Comments as predictor fixed effects, and allowing participants' intercepts to vary (see Tables 16-25). Consistent with the hypotheses, there were significant unique effects of PMAPS (*Prejudice of Post: F*(1, 194) = 29.52, $p < .001$; *Prejudice of Person: F*(1, 194) = 41.45, $p < .001$, *Righteousness of Commenter F*(1, 194) = 4.55, $p = .036$, *Positive Interaction: F*(1, 194) = 6.42, $p = .013$, *Negative Interaction: F*(1, 194) = 21.14, $p < .001$; see PMAPS β values in Tables 16-25) such that, generally people higher in PMAPS viewed posts and people as both more prejudiced and more malicious, viewed commenters as doing the right thing, and were less likely to interact positively and more likely to interact negatively with the posts. Also, there were significant unique effects of Need for Chaos (*Maliciousness of Post*: *F*(1, 194) = 18.07, $p < .001$, *Maliciousness of Person F*(1, 194) = 12.61, $p < .001$; see Need for Chaos β values in Tables 16-25) such that, generally people higher in Need for Chaos viewed posts and people as less malicious, perhaps because need for chaos is associated with sympathizing with online behavior that bucks social norms. There were additional significant unique effects of SV (*Righteousness of Commenter*: *F*(1, 194) = 7.73, $p = .007$, *Positive Interaction: F*(1, 194) = 19.41, $p < .001$, *Negative Interaction: F*(1, 194) = 4.95, $p = .029$; see SV β values in Tables 16-25) such that people higher in SV rated the confronting commenter as generally doing more of the right thing and were more likely to interact with the

posts themselves, whether positively or negatively. This makes sense because SV has not been shown to be associated with a particular viewpoint (that is, SV is unrelated to racist views), but is associated with wanting to impress one's own opinions onto others. Therefore, it follows that SV is associated with posting one's own opinions and commenting online, regardless of the views expressed in those interactions.

Additionally, consistent with the Trolling hypothesis, there were significant unique effects of Anonymity (*Prejudice of Person: F*(1, 530) = 3.19, *p* = .045, *Maliciousness of Person F*(1, 530) = 4.12, *p* = .018, *Honesty of Post: F*(1, 530) = 3.88, *p* = .049, *Honesty of Person: F*(1, 530) = 4.87, *p* = .008, *Righteousness of Commenter: F*(1, 530) = 4.24, *p* = .040, *Negative Interaction: F*(1, 194) = 30.82, *p* < .001; see Anonymity β values in Tables 16-25), such that people posting anonymous posts were rated as less prejudiced, malicious, and honest than people posting identifiably, and such that identifiable posts were more likely to be negatively interacted with. This could suggest that people view anonymous posts as trolling, not meant to be taken seriously or as truth, but rather intended to garner extreme reactions by bucking against the social order. There were also significant unique effects of presence of Comments (*Prejudice of Post: F*(1, 530) = 3.94, *p* = .047; *Maliciousness of Post*: *F*(1, 530) = 9.54, *p* < .001, *Prejudice of Person: F*(1, 530) = 3.57, *p* = .034, *Maliciousness of Person F*(1, 530) = 3.61, *p* = .026; , *Negative Interaction: F*(1, 194) = 13.03, *p* < .001; see Comments β values in Tables 16-25) such that posts with confrontational comments were rated as generally more malicious and more prejudiced and were more likely to be negatively interacted with. This may be because the presence of a confronting comment indicates that another person believed the post was malicious and prejudiced enough to warrant an argument. These main effects were qualified by significant two-way interactions between Anonymity and presence of Comments (*Honesty of Post: F*(1,

530) = 3.92, p = .042, *Honesty of Person:* $F(1, 530) = 3.16$, p = .047, *Attention Seeking*: $F(1, 530) = 4.57$, p = .010, *Negative Interaction:* $F(1, 194) = 10.08$, p = .006; see interaction term β values in Tables 16-25). These interactions indicate that the effects of anonymity of post on these criterion variables depended upon whether the post included confrontational arguments from others.

I then conducted simple slopes analyses on the interaction terms that were significant to determine whether the final hypotheses were supported (see Table 26). As predicted, identifiable posts that included confrontational comments were rated as more honest and less attention-seeking than anonymous posts regardless of the presence of commenters. Such posts were also more likely to provoke negative actions (e.g., blocking, leaving a negative comment). This is consistent with the Trolling Hypothesis, which predicted that anonymous posts would be seen as less honest and more attention seeking and that comments on such anonymous posts would be seen as evidence of "feeding the trolls." It is also possible that the presence of comments on Facebook posts (the identifiable condition in this study) are more powerful indicators of an OP's honest levels of prejudice because the people commenting on a Facebook post might be assumed to know the OP in real life.

Additionally, to test the viability of the Lawless Model of Perceiving Trolling, a path analysis was conducted on the parts of the model relevant to Study 2 (see Figure 4). Consistent with past research (Lawless & Saucier a), with Study 1, and with the hypotheses, the perceived prejudice of a post and the anonymity of the Original Person (OP) who posted it influence third-party perceptions of trolling. The perception of trolling is then negatively associated with the likelihood of confronting the post via leaving a negative interaction. Additionally, the third-party's individual PMAPS score was shown to be positively associated with perceiving posts as

prejudiced and with confronting racist posts regardless of perceived trolling. Additionally, SV was positively associated with confronting posts regardless of perceptions of trolling, perhaps because of SV's association with counterarguing and the impression of one's beliefs onto others. These new additions to the model can now be integrated and contextualized with the findings of past research.

Chapter 4 - General Discussion

While past studies have focused on how anonymity of both the perpetrator and target affects third-party perceptions of racial prejudice online (Lawless & Saucier, submitted a; submitted b; in preparation), the current studies have helped clarify how individual differences within the third-party affect these perceptions. They demonstrate that individual differences, including social vigilantism and personal endorsement of prejudicial beliefs, affect how third parties perceive and choose to interact with prejudicial statements online. These studies contribute to the existing literature on online behavior and elucidate some of the potential similarities and differences between online and traditional face-to-face interaction.

Study 1 examined the attributions and judgements people make in response to confrontations of online prejudice in both anonymous and identifiable conditions. Several mock posts displaying overt racism were created and placed those posts in either an anonymous or identifiable social media environment similar to Lawless and Saucier (submitted a; submitted b; in preparation). Additionally, these posts were either presented with no other interactions or with a single other commenter who confronts the Original Person who posted (the OP). Participants also filled out individual difference measures of SV, PMAPS, and ATB, and rated how racist they found the post, how racist they found the OP, to what extent they believed OP agrees with the sentiments expressed, and to what extent they believed the post was made to garner attention or disturb the status quo. These items were based on Lawless and Saucier (submitted a; submitted b). Study 2 used the same methods as Study 1, but this time participants were asked how they would interact with the posts and gave them a chance to write their own comment on the posts. Therefore, participants in Study 2 were evaluating interactions as potential interactors rather than as third-parties.

Previous research has found that, though SV is not associated with being well-informed, it is associated with the use of strategies to prevent persuasion, including counterarguing (Saucier et al., 2021) and that SV is associated with counterarguing in response to extreme political opinions (e.g., Saucier & Webster, 2010; Raimi & Leary, 2014). Because SV is related to counterarguing in this way, it makes sense that SV would be associated with viewing confronting prejudice online as a positive behavior. Consistent with this previous research and with the hypotheses, in Study 1, SV was positively associated with saying that the person confronting the comments did the right thing. Similarly, in Study 2, SV was associated with leaving more confronting comments. This is consistent with the aforementioned past findings that SV is associated with counterarguing.

Another individual difference that may affect views of arguing against prejudice online are participants' own prejudicial attitudes. Miller and Saucier (2018) created the PMAPS to assess the tendency to attribute ambiguous situations and actions to prejudice. The PMAPS has been shown to predict attributions to prejudice in a variety of situations (e.g., Miller et al., 2017; Stratmoen et al., 2019). Because extremely prejudiced statements online are ambiguous and can be attributed to either genuine prejudice or to trolling behaviors, it makes sense that PMAPS may affect views of such statements. Consistent with this past research and the hypotheses, in Study 1, PMAPS was associated with perceptions of the mock posts such that participants who displayed more prejudicial beliefs rated prejudicial statements as less racist and more honest. Racist people are likely to see racism as less problematic and may rate even overtly racist statements as less racist because of this. It is also possible that racist people may see racism as more common because they identify with racists, and they therefore may rate racist posts as more honest. Study 2 additionally found that PMAPS was associated with comments such that anti-

prejudiced attitudes were associated with more confrontation of prejudice via negatively reacting to posts (e.g., blocking the OP, leaving a negative comment). This is also consistent with past research (e.g., Czopp et al., 2006) and it makes intuitive sense that people with more anti-racist attitudes would more readily confront racism.

These studies explore how these individual differences and anonymity affect third-party views of confrontations to prejudice. Some previous research has suggested online anonymity is negatively associated with perceived honesty (e.g., Lawless & Saucier, submitted a; submitted b; Reader, 2012). This is consistent with the Trolling Hypothesis, wherein people tend to think of online anonymity as a cover for groundless aggression. In contrast, some previous findings have suggested online anonymity is positively associated with perceived honesty (e.g., Sticca & Perren, 2012), which is consistent with the Disinhibition Hypothesis, wherein people tend to think of online anonymity as an instrument people use to protect themselves from possible social consequences of advocating their own genuine beliefs. While the effects of pure anonymity on these perceptions have been documented (e.g., Lawless & Saucier, submitted a), it is likely that including a confrontational comment may change these perceptions. It is possible that such an argument will make participants see the OP as more honest, or perhaps as more of an attention-seeker who simply likes to argue (i.e., a troll).

Across Studies 1 and 2, consistent with previous research by Lawless and Saucier (submitted b) and with the hypotheses, there were negative correlations between perceived honesty of the OP and perceived attention seeking of the OP. This pattern of low honesty and high attention seeking is consistent with common conceptions of trolling behavior. Furthermore, it was found that perceived righteousness of the commenter was positively correlated with perceived honesty of the post and OP, which is consistent with the idea that arguing with posts

perceived as trolling is seen as the wrong thing to do, as "feeding the trolls" and reinforcing the trolling behavior. It also makes sense that similar results were found across both studies concerning items measuring perceptions of the OP since those perceptions may not be affected by the opportunity to participate.

The Trolling Hypothesis states that anonymous statements will be rated as less honest and prejudiced, and more attention-seeking than identifiable statements because they are thought of as "trolls" who are trying to create chaos and discord rather than espousing genuine beliefs. In contrast, the second hypothesis, the Disinhibition Hypothesis, states that anonymous statements will be rated as more honest and prejudiced, and less attention-seeking than identifiable statements because anonymity is thought of as removing the social pressures that usually inhibit genuine hate speech. In Study 1, confrontational comments on anonymous posts were generally perceived negatively. This is consistent with the Trolling Hypothesis because the comments may be seen as "feeding the trolls." This is not consistent with the Disinhibition Hypothesis, where confrontational comments would be perceived positively because they are perceived as arguing against honest prejudice. Similarly, Study 2 found that posts that are identifiable received more interactions (i.e., negative Facebook reactions, comments), which may indicate that they are generally seen as more honest, which is consistent with the Trolling Hypothesis because interacting with anonymous posts may be seen as "feeding the trolls," but interacting with identifiable posts may be seen as trying to confront genuine prejudice. This is not consistent with the Disinhibition Hypothesis, which would lead to the hypothesis that anonymous posts should receive more interactions because participants may be emboldened by their own anonymity, Similarly, posts that already had confronting comments on them were interacted with more than posts that had no comments. This may be because participants may feel more comfortable taking

an action someone else already clearly agrees with. The presence of a comment, especially on Facebook where it may be presumed that the commenter knows the OP, may indicate that others believe the OP really is prejudiced and is not just "trolling."

My past research findings on trolling in social media (Lawless & Saucier, submitted a; submitted b; in preparation) can be expressed by the Lawless Model of Perceiving Trolling (see Figure 1). Essentially, the perceived prejudice of a post, the anonymity of the Original Person (OP) who posted it, and the target of the post being a group of people influence third-party perceptions of trolling (i.e., decreased perceptions that the post reflects what the OP really believes, increased perceptions of the OP as attention-seeking, and higher ratings of the OP as a "troll"). Additionally, the third-party's individual PMAPS score has also been shown to be positively associated with perceiving posts as prejudiced as well as perceiving the OP as a troll. The current research further tested this model as well as extended it.

The current research added a new outcome variable, confrontation of the post by the third party, to the model (see Figure 2). Consistent with the hypotheses, perceived trolling was generally negatively associated with confrontation of the post. This would make sense because perceived trolling indicates that the third party does not think the OP truly believes the prejudiced ideas contained in the post. If the OP does not really believe what they are saying, there may be less reason to confront them. However, PMAPS scores were positively associated with confrontation of the post regardless of perceived trolling. PMAPS is associated with perceiving prejudice more readily and with confronting it more often. Therefore, it makes sense that PMAPS is positively associated with confronting prejudiced posts regardless of the perceived intention of the OP. Additionally, the current research added a new individual difference predictor to the model, Social Vigilantism (SV). Again, consistent with the

hypotheses, SV was also positively associated with confrontation of prejudiced posts regardless of perceptions of trolling. SV indicates a desire to impress one's own beliefs onto others and is associated with counterarguing and other strategies that indicate a willingness to confront statements one disagrees with. Given that prejudice is generally not socially acceptable, it is reasonable to expect that most participants will disagree with the content of extremely prejudiced posts. Therefore, SV should be associated with confronting those posts. Given the findings of the current studies, the full Lawless Model of Perceiving Trolling can now be expressed as in Figure 5. This full model combines the model backed by previous research with the new information added by the current two studies.

The potential implications of the current studies are that factors of the online environment, such as anonymity of platform, as well as individual differences affect how individuals react to prejudicial speech online. Theoretical implications include adding further evidence in the debate between the Trolling and Disinhibition Hypotheses of online behavior, generally in favor of the Trolling Hypothesis. The practical implications of these studies include insights into how prejudice might be confronted online as well as what factors may contribute to getting social media consumers to behave in a way that is actively anti-prejudice. These and future studies along this line of research are important to fully understand the factors at play within internet culture.

Limitations and Future Directions

The current studies are not without limitations. The first limitation is the cross-sectional nature of the current studies. This limits the ability to draw causal conclusions about the relationships between current political climate, participants' levels of racial prejudice, and their perceptions of online racial hate speech. One could make the argument that participants' levels

of racial prejudice are fairly consistent across time. However, researchers should be hesitant to draw concrete causal conclusions from the proposed studies, particularly given the recent uptick in publicized hate speech in the United States' current political climate (Waltman, 2018). It is possible that this recent lift of prejudice suppression might lead participants to believe that online hate speech is more honest than they would in a different climate because public racial hate speech has become more salient in the news over the past few years. It is also possible that, because the majority of the participants in the current studies were White, the conclusions are skewed toward a White perspective of the relationship between racism and trolling. For example, is it possible that participants of Color find racist speech easier to perceive as honest because they are more likely to encounter it in the real world. Steps should be taken in the future to extend this research by examining perceptions of online behavior under different political and social circumstances.

Another limitation in the current studies is the usage of mock posts that were free of other common factors (e.g., name of subreddit). In conducting studies in this fashion, participants are not able to perceive some online social cues and other indications surrounding the intent of the perpetrator of the racism. For example, some subreddits cater to groups dedicated to pointing out racism which might justify a racist post and make the mock Reddit posts without subreddit identifiers more ambiguous than the mock Facebook posts. That said, there are ethical concerns in the employment of more realistic procedures (e.g., using posts actually found online). These concerns include the issue of doxing, wherein a participant could google the posts to find a real person's social media account as well as issues with anonymity, including being unable to ensure anonymous posts were not made by children. Additionally, the inclusion of comments on Facebook may carry more weight than those on Reddit because Facebook reactions are expected

to come from a person's friends while Reddit reactions come from strangers. Third party observers may react differently to community reactions coming from someone's inner social circle than they would to the reactions of people who do not know the person posting. Thus, the results of the current studies may not generalize to real world events. Future studies should add more realistic online interactions and could manipulate the types of communities the posts come from by manipulating subreddit or Facebook group names or making it clear that the Facebook posts were made on public forums rather than someone's personal timeline. Future studies could also ask participants whether they would prefer to comment anonymously or identifiably to examine the conditions which might incite users to prefer anonymity online.

Another aspect not addressed by the current studies is that online anonymity is not always accompanied by negative expressions of prejudice. Online anonymity is neither good nor bad, and, while it may be used as a tool to express vile beliefs without social consequence, it may also be used to safely express oneself in a positive way. It may allow people to safely discuss their sexual orientation in places where doing so would otherwise be dangerous, or allow people to discuss negative aspects of totalitarian governments. The current studies were focused solely on negative uses of the online environment, but future studies should include positive uses to assess public responses to both identifiable and anonymous positive messages. This research could lead to valuable contributions about how best to spread positive messages as well as how best to tamp down violent or prejudicial messages.

Conclusion

Across two studies, the effects of anonymity, confrontational comments, and individual differences on third-party perceptions of the honesty and racism of online prejudiced speech as well as the third-parties' estimates of how they would interact with the posts themselves were

examined. These studies are timely and extend the existing literature on internet behavior by further investigating the relationships between various online social conditions and community reactions based on those conditions. The potential implications of the current studies are that factors of the online environment, such as anonymity of platform and initial community reactions affect how people interact with online hate-speech. Specifically, consistent with the Trolling Hypothesis, it is possible that, on identifiable platforms, people may use community reactions to judge the people who post such rhetoric while such speech gets more of a "pass" on anonymous platforms due to the possibility of trolling. These studies demonstrate that many people use online group norms to decide what behavior is acceptable rather than using the norms that exist in face-to-face interactions. These studies also show that, in groups where it is implied that prejudice may be the norm, people may not take online prejudiced rhetoric seriously, which could foster toxic online environments that are conducive to cyberbullying and even incitements to real-world violence against marginalized groups. In such an environment, users may even dismiss legitimate threats of violence as trolling and ignore important warnings of impending incidents. Certain kinds of trolling, such as the spread of misinformation may be extra insidious because of the sleeper effect, wherein information is later remembered as true even when it is recognized as false trolling at the moment of presentation (e.g., Andriopoulos, 2011; Petrocelli, Seta, & Seta, 2023). While there is research that examines how to train human users to identify misinformation, much of it has a poor success rate thus far, and there has been very little research on the identification of trolling (e.g., Jones-Jang, Mortensen, & Liu, 2019; McCright & Dunlap, 2017). Therefore, these and future studies along this line of research are important to fully understand the factors that contribute to internet culture. While the internet has a unique

ability to bring people together in truly global communities, it also may have the potential to foster sinister communities based on deindividuated hatred.

Chapter 5 - Tables

Table 1

Means, Standard Deviations, and Bivariate Correlations between Predictor Variables in Study 1

Variable	*M (SD)*	1	2	3	4	5
1. Social Desirability	13.72 (5.78)	(.87)				
2. ATB	3.22 (1.32)	-.57***	(.90)			
3. PMAPS	6.36 (1.36)	.10	-.62***	(.88)		
4. Need for Chaos	4.04 (2.27)	-.36**	.30**	-.27**	(.85)	
5. Social Vigilantism	5.78 (1.64)	.12	-.14	.13	.16	(.90)

*$p \leq .05$, **$p \leq .01$, ***$p \leq .001$

Table 2

Bivariate Correlations between Predictor Variables and Participant Demographics in Study 1

Variable	Participant Gender	Participant Ethnicity
1. Social Desirability	.13	.08
2. ATB	-.15	-.27*
3. PMAPS	.10	.31*
4. Need for Chaos	-.16	.10
5. Social Vigilantism	.12	-.14

*$p \leq .05$, **$p \leq .01$, ***$p \leq .001$

Note: For the purposes of these correlations, men were coded as 1 and women were coded as 2. White participants were coded as 1 and participants of Color were coded as 2.

Table 3

Means, Standard Deviations, and Bivariate Correlations between Criterion Variables in Study 1

Variable	*M (SD)*	1	2	3	4	5	6	7	8
1. Post Prejudiced	8.10 (1.50)	-							
2. Post Malicious	7.25 (2.07)	.66***	-						
3. Post Honesty	4.97 (2.19)	.29**	.30**	-					
4. Person Prejudiced	5.87 (1.69)	.80***	.71***	.35**	-				
5. Person Malicious	4.22 (2.10)	.64**	.87***	.33**	.72***	-			
6. Person Honesty	4.27 (1.74)	.29**	.31**	.46***	.35**	.33**	-		
7. Attention Seeking	4.04 (2.22)	.27**	.35**	-.31**	.37**	.42**	-.21*	-	
8. Commenter Right	6.36 (2.19)	.14	.16	.34**	.31**	.20	.28**	-.21*	-

*$p \leq .05$, **$p \leq .01$, ***$p \leq .001$

Table 4

Fixed Effects Summary Table for Perceptions of the Post as Prejudiced in Study 1

Predictor Variable	*B*	*SE*	*F*	*p*
PMAPS	0.56	0.09	6.06	<.001
Social Vigilantism (SV)	-0.11	0.08	1.38	.172
Need for Chaos	-0.06	0.05	1.56	.210
Anonymity	0.04	0.04	1.19	.233
Inclusion of Comments	0.14	0.04	3.46	<.001
Anonymity*Comments	-0.01	0.04	0.01	.999
SV*Comments	-0.02	0.02	0.93	.354
Anonymity*SV*Comments	-0.03	0.02	1.52	.129

Table 5

Fixed Effects Summary Table for Perceptions of the Post as Malicious in Study 1

Predictor Variable	*B*	*SE*	*F*	*p*
PMAPS	0.80	0.11	7.12	<.001
Social Vigilantism (SV)	-0.07	0.09	0.80	.428
Need for Chaos	-0.29	0.06	4.81	<.001
Anonymity	0.02	0.05	0.34	.734
Inclusion of Comments	0.23	0.05	4.52	<.001
Anonymity*Comments	0.02	0.05	0.34	.735
SV*Comments	-0.05	0.03	1.42	.155
Anonymity*SV*Comments	-0.04	0.03	1.32	.189

Table 6

Fixed Effects Summary Table for Perceptions of the Post as Honest in Study 1

Predictor Variable	*B*	*SE*	*F*	*p*
PMAPS	-0.36	0.14	2.60	.011
Social Vigilantism (SV)	0.25	0.11	2.15	.035
Need for Chaos	-0.06	0.05	1.56	.210
Anonymity	-0.18	0.05	3.57	.002
Inclusion of Comments	-0.01	0.05	0.16	.877
Anonymity*Comments	-0.15	0.05	2.97	.003
SV*Comments	-0.01	0.03	0.27	.790
Anonymity*SV*Comments	0.02	0.03	0.67	.502

Table 7

Fixed Effects Summary Table for Perceptions of the Person as Prejudiced in Study 1

Predictor Variable	*B*	*SE*	*F*	*p*
PMAPS	0.75	0.09	8.70	<.001
Social Vigilantism (SV)	-0.02	0.07	0.23	.820
Need for Chaos	-0.06	0.06	1.03	.490
Anonymity	-0.25	0.04	6.24	.004
Inclusion of Comments	0.14	0.04	3.53	<.001
Anonymity*Comments	0.01	0.04	0.07	.941
SV*Comments	0.01	0.03	0.36	.723
Anonymity*SV*Comments	-0.03	0.02	1.16	.245

Table 8

Fixed Effects Summary Table for Perceptions of the Person as Malicious in Study 1

Predictor Variable	*B*	*SE*	*F*	*p*
PMAPS	0.82	0.12	7.05	<.001
Social Vigilantism (SV)	-0.02	0.10	0.16	.870
Need for Chaos	-0.47	0.13	3.62	<.001
Anonymity	-0.28	0.05	5.64	.001
Inclusion of Comments	0.23	0.05	4.52	<.001
Anonymity*Comments	0.07	0.05	1.38	.167
SV*Comments	-0.06	0.03	1.89	.059
Anonymity*SV*Comments	-0.03	0.03	0.99	.325

Table 9

Fixed Effects Summary Table for Perceptions of the Person as Honest in Study 1

Predictor Variable	*B*	*SE*	*F*	*p*
PMAPS	-0.56	0.10	5.83	<.001
Social Vigilantism (SV)	0.07	0.08	0.84	.403
Need for Chaos	-0.06	0.05	1.56	.210
Anonymity	-0.22	0.05	4.36	.018
Inclusion of Comments	0.02	0.05	0.47	.638
Anonymity*Comments	-0.23	0.05	4.54	.006
SV*Comments	-0.04	0.03	1.38	.118
Anonymity*SV*Comments	0.01	0.03	0.23	.818

Table 10

Fixed Effects Summary Table for Perceptions of Attention Seeking in Study 1

Predictor Variable	*B*	*SE*	*F*	*p*
PMAPS	0.46	0.16	2.97	.004
Social Vigilantism (SV)	0.11	0.13	0.83	.411
Need for Chaos	-0.06	0.05	1.56	.210
Anonymity	-0.03	0.05	0.59	.553
Inclusion of Comments	0.01	0.05	0.16	.876
Anonymity*Comments	0.07	0.03	2.20	.029
SV*Comments	-0.01	0.03	0.22	.824
Anonymity*SV*Comments	-0.01	0.05	0.05	.996

Table 11

Fixed Effects Summary Table for Perceptions of Righteousness of Commenter in Study 1

Predictor Variable	*B*	*SE*	*F*	*p*
PMAPS	0.48	0.16	2.99	.002
Social Vigilantism (SV)	0.60	0.13	4.63	<.001
Need for Chaos	-0.04	0.05	0.87	.268
Anonymity	-0.13	0.05	2.54	.018
Anonymity*Comments	0.01	0.05	0.30	.762
SV*Comments	-0.01	0.03	0.22	.824
Anonymity*SV*Comments	-0.03	0.03	1.19	.290

Table 12

Results from Simple Slopes Analyses of Significant Interactions

	Anonymity		*Presence of Comments*	
Perception	*r*	*t*	*r*	*t*
Honesty of Post	-0.37	-2.82*	0.41	3.21*
Honesty of Person	-0.35	-2.65*	0.38	2.98*
Attention Seeking	0.28	2.07*	0.29	2.08*

*$p \leq .05$

Table 13

Means, Standard Deviations, and Bivariate Correlations between Predictor Variables in Study 2

Variable	*M (SD)*	1	2	3	4	5
1. Social Desirability	14.21 (5.36)	(.82)				
2. ATB	2.96 (1.29)	-.46**	(.79)			
3. PMAPS	6.44 (1.41)	.13	-.68***	(.91)		
4. Need for Chaos	3.97 (2.18)	-.43**	.34**	-.31*	(.94)	
5. Social Vigilantism	5.7 (1.39)	.12	-.09	.13	-.10	(.89)

*$p \leq .05$, **$p \leq .01$, ***$p \leq .001$

Table 14

Bivariate Correlations between Predictor Variables and Participant Demographics in Study 2

Variable	Participant Gender	Participant Ethnicity
1. Social Desirability	.18	.04
2. ATB	-.14	-.29*
3. PMAPS	.25*	.33*
4. Need for Chaos	-.13	.14
5. Social Vigilantism	.17	-.16

*$p \leq .05$, **$p \leq .01$, ***$p \leq .001$

Note: For the purposes of these correlations, men were coded as 1 and women were coded as 2. White participants were coded as 1 and participants of Color were coded as 2.

Table 15

Means, Standard Deviations, and Bivariate Correlations between Criterion Variables in Study 2

Variable	*M* (*SD*)	1	2	3	4	5	6	7	8	9	10
1. Post Prejudiced	8.09 (1.53)	-									
2. Post Malicious	7.44 (1.99)	.63***	-								
3. Post Honesty	4.23 (2.14)	.43***	.35***	-							
4. Person Prejudiced	7.95 (1.56)	.83***	.63***	.44***	-						
5. Person Malicious	7.46 (1.93)	.65***	.81***	.36***	.67***	-					
6. Person Honesty	4.89 (1.56)	.69***	.53***	.48***	.69***	.56***	-				
7. Attention Seeking	6.17 (2.00)	.35***	.28***	-.12**	.30***	.28***	-.19***	-			
8. Commenter Right	5.63 (2.07)	.21***	.24***	.15**	.26***	.33***	.18***	-.20***	-		
9. Pos. Interaction	2.26 (2.40)	-.44***	-.23***	-.16***	-.40***	-.25***	-.37***	-.08*	-.17***	-	
10. Neg Interaction	6.33 (3.07)	.23***	.30***	.18***	.30***	.35***	.21***	.06	.34***	-.03	-

*$p \leq .05$, **$p \leq .01$, ***$p \leq .001$

Table 16

Fixed Effects Summary Table for Perceptions of the Post as Prejudiced Study 2

Predictor Variable	*B*	*SE*	*F*	*p*
PMAPS	0.50	0.09	29.52	<.001
Social Vigilantism (SV)	-0.12	0.09	1.58	.213
Need for Chaos	-0.04	0.09	0.87	.268
Anonymity	0.06	0.03	3.75	.053
Inclusion of Comments	0.07	0.03	3.94	.047
Anonymity*Comments	0.02	0.03	0.29	.590
SV*Comments	0.01	0.02	0.23	.629
Anonymity*SV*Comments	0.01	0.02	0.02	.877

Table 17

Fixed Effects Summary Table for Perceptions of the Post as Malicious Study 2

Predictor Variable	*B*	*SE*	*F*	*p*
PMAPS	0.63	0.11	31.97	<.001
Social Vigilantism (SV)	-0.07	0.11	0.34	.564
Need for Chaos	-0.16	0.09	18.07	<.001
Anonymity	0.02	0.05	0.11	.743
Inclusion of Comments	0.11	0.05	9.54	<.001
Anonymity*Comments	-0.01	0.05	0.02	.890
SV*Comments	0.03	0.04	0.56	.456
Anonymity*SV*Comments	-0.02	0.04	0.30	.584

Table 18

Fixed Effects Summary Table for Perceptions of the Post as Honest Study 2

Predictor Variable	*B*	*SE*	*F*	*p*
PMAPS	-0.31	0.14	5.29	.024
Social Vigilantism (SV)	-0.16	0.14	1.38	.244
Need for Chaos	-0.04	0.05	0.11	.856
Anonymity	-0.08	0.04	3.88	.049
Inclusion of Comments	-0.05	0.06	0.78	.379
Anonymity*Comments	-0.13	0.06	3.92	.042
SV*Comments	-0.02	0.04	0.18	.670
Anonymity*SV*Comments	0.04	0.06	0.56	.456

Table 19

Fixed Effects Summary Table for Perceptions of the Person as Prejudiced Study 2

Predictor Variable	*B*	*SE*	*F*	*p*
PMAPS	0.57	0.09	41.45	<.001
Social Vigilantism (SV)	-0.11	0.09	1.42	.237
Need for Chaos	-0.04	0.03	0.11	.268
Anonymity	-0.08	0.03	3.19	.045
Inclusion of Comments	-0.09	0.03	3.57	.034
Anonymity*Comments	0.02	0.03	0.25	.614
SV*Comments	-0.01	0.03	0.07	.793
Anonymity*SV*Comments	0.03	0.03	1.13	.288

Table 20

Fixed Effects Summary Table for Perceptions of the Person as Malicious Study 2

Predictor Variable	*B*	*SE*	*F*	*p*
PMAPS	0.63	0.11	33.23	<.001
Social Vigilantism (SV)	-0.09	0.11	0.70	.406
Need for Chaos	-0.20	0.09	12.61	<.001
Anonymity	-0.20	0.05	4.12	.018
Inclusion of Comments	-0.17	0.05	3.61	.026
Anonymity*Comments	0.03	0.05	0.23	.634
SV*Comments	0.02	0.03	0.34	.559
Anonymity*SV*Comments	-0.01	0.03	0.03	.864

Table 21

Fixed Effects Summary Table for Perceptions of the Person as Honest Study 2

Predictor Variable	*B*	*SE*	*F*	*p*
PMAPS	-0.51	0.09	35.49	<.001
Social Vigilantism (SV)	-0.10	0.09	1.36	.247
Need for Chaos	-0.04	0.05	0.11	.842
Anonymity	-0.16	0.04	4.87	.008
Inclusion of Comments	-0.01	0.04	0.02	.903
Anonymity*Comments	0.09	0.04	3.16	.047
SV*Comments	0.01	0.03	0.08	.782
Anonymity*SV*Comments	0.05	0.03	2.99	.085

Table 22

Fixed Effects Summary Table for Perceptions of Attention Seeking Study 2

Predictor Variable	*B*	*SE*	*F*	*p*
PMAPS	0.20	0.12	2.76	.101
Social Vigilantism (SV)	-0.02	0.12	0.04	.841
Need for Chaos	-0.04	0.06	0.99	.324
Anonymity	-0.06	0.06	0.99	.320
Inclusion of Comments	-0.06	0.06	1.01	.315
Anonymity*Comments	0.19	0.06	4.57	.010
SV*Comments	-0.02	0.04	0.17	.680
Anonymity*SV*Comments	0.02	0.04	0.23	.629

Table 23

Fixed Effects Summary Table for Perceptions of Righteousness of Commenter Study 2

Predictor Variable	*B*	*SE*	*F*	*p*
PMAPS	0.44	0.12	4.55	.036
Social Vigilantism (SV)	0.59	0.21	7.73	.007
Need for Chaos	-0.04	0.21	0.09	.868
Anonymity	-0.10	0.05	4.24	.040
Anonymity*Comments	-0.03	0.06	0.26	.611
SV*Comments	-0.06	0.05	2.01	.157
Anonymity*SV*Comments	-0.01	0.06	0.01	.953

Table 24

Fixed Effects Summary Table for Self-Reported Likelihood of Leaving a Positive Interaction

Predictor Variable	*B*	*SE*	*F*	*p*
PMAPS	-0.39	0.16	6.42	.013
Social Vigilantism (SV)	0.70	0.16	19.41	<.001
Need for Chaos	-0.04	0.06	0.82	.368
Anonymity	-0.04	0.04	0.82	.366
Inclusion of Comments	-0.04	0.04	0.76	.383
Anonymity*Comments	0.03	0.04	0.51	.474
SV*Comments	-0.04	0.03	1.46	.228
Anonymity*SV*Comments	0.02	0.03	0.29	.589

Table 25

Fixed Effects Summary Table for Self-Reported Likelihood of Leaving a Negative Interaction

Predictor Variable	*B*	*SE*	*F*	*p*
PMAPS	0.84	0.18	21.14	<.001
Social Vigilantism (SV)	0.41	0.18	4.95	.029
Need for Chaos	-0.04	0.06	0.99	.287
Anonymity	-0.38	0.07	30.82	<.001
Inclusion of Comments	-0.16	0.07	13.03	<.001
Anonymity*Comments	0.15	0.07	10.08	.006
SV*Comments	-0.03	0.05	0.38	.536
Anonymity*SV*Comments	0.01	0.05	0.07	.789

Table 26

Results from Simple Slopes Analyses of Significant Interactions Between Anonymity and Presence of Comments in Study 2

	Anonymity		*Presence of Comments*	
Perception	*r*	*t*	*r*	*t*
Honesty of Post	-0.29	-2.08*	0.38	2.97*
Honesty of Person	-0.31	-2.68*	0.36	2.82*
Attention Seeking	0.26	2.01*	0.27	2.05*
Negative Interaction	-0.38	-3.17*	0.47	4.07*

*$p \leq .05$

Chapter 6 - Figures

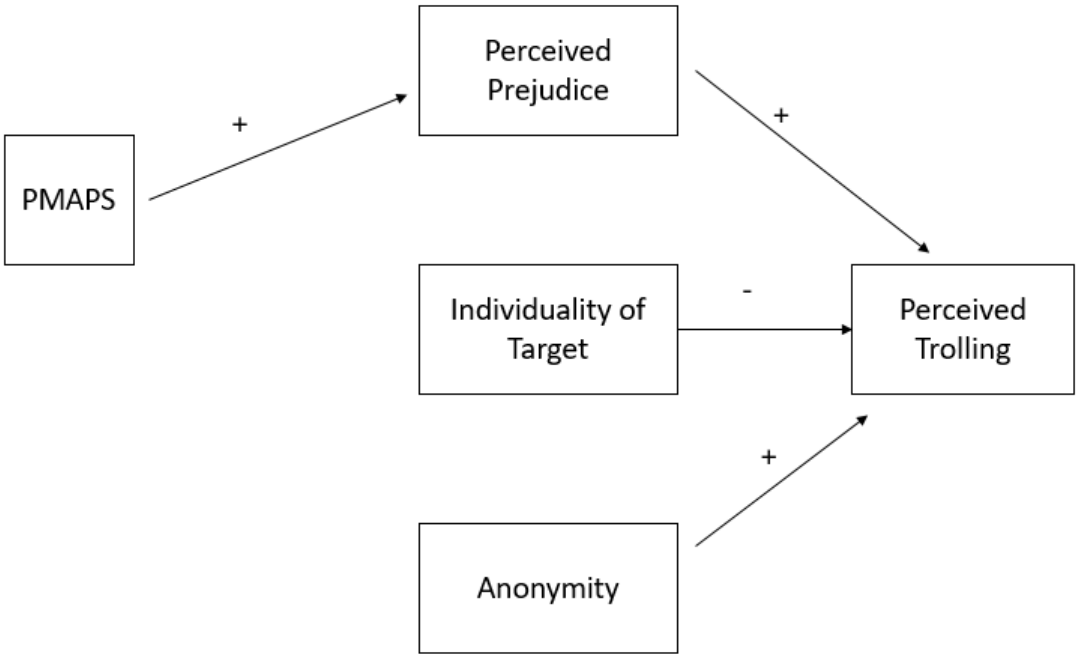

Figure 1: Lawless Model of Perceiving Trolling According to Past Research
Note: "Perceived Trolling" here and in future figures refers to the combination of perceiving posts and OPs as less honest and more attention seeking

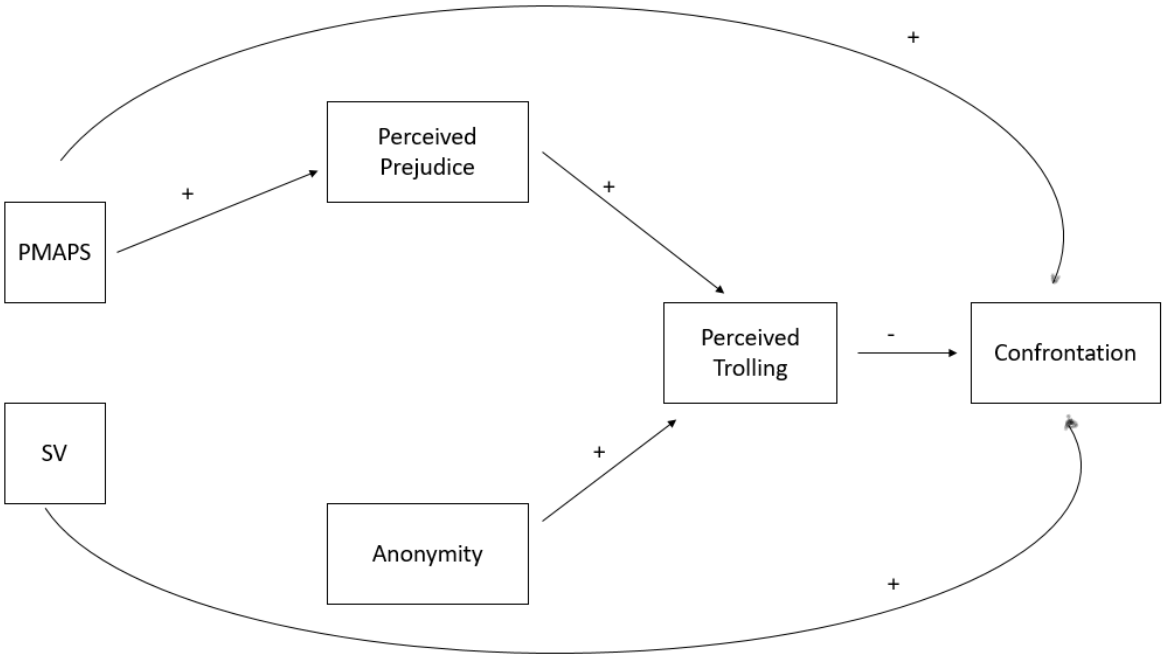

Figure 2: Lawless Model of Perceiving Trolling Being Tested in Current Studies

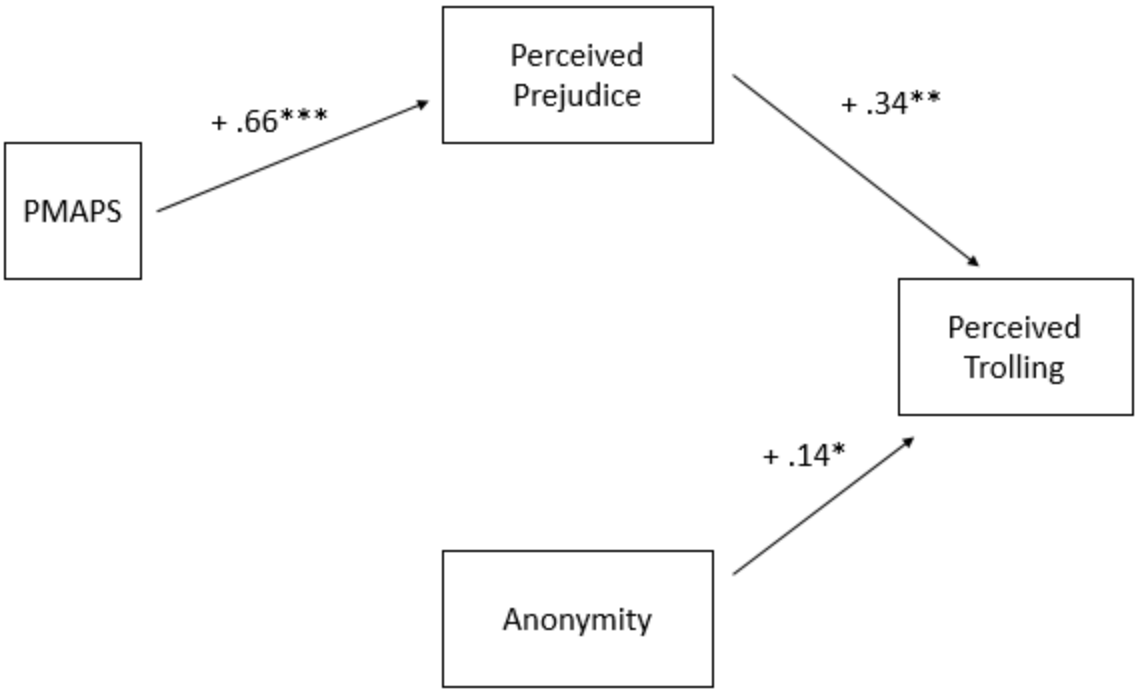

Figure 3: Lawless Model of Perceiving Trolling as Tested in Study 1
*$p \leq .05$, **$p \leq .01$, ***$p \leq .001$

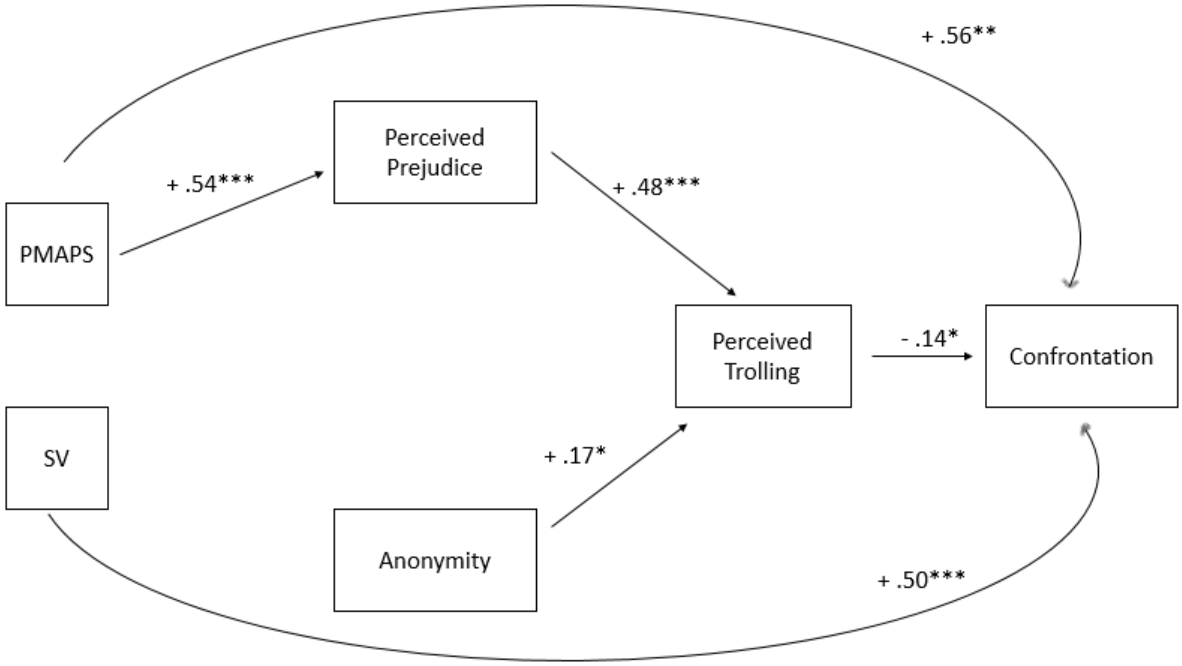

Figure 4: Lawless Model of Perceiving Trolling as Tested in Study 2
*$p \leq .05$, **$p \leq .01$, ***$p \leq .001$

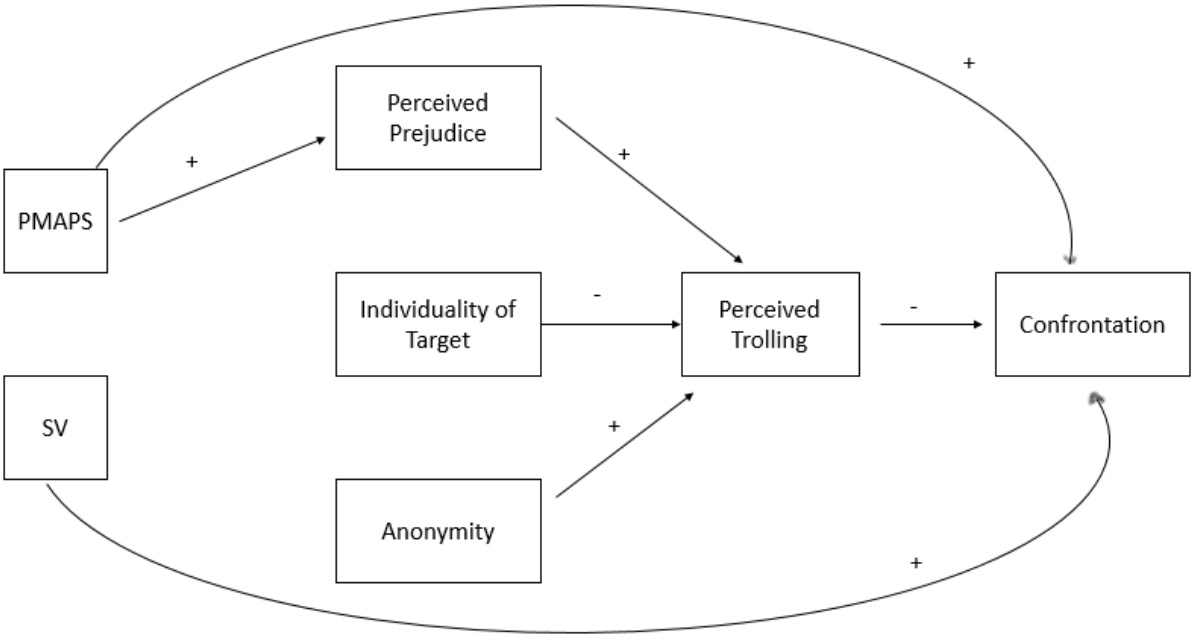

Figure 5: Lawless Model of Perceiving Trolling in Full

References

Agadjanian, A., Carey, J. M., Horiuchi, Y., & Ryan, T. J. (2021). Disfavor or Favor? Assessing the Valence of White Americans' Racial Attitudes. *Quarterly Journal of Political Science*.

Anderson, J., Bresnahan, M., & Musatics, C. (2014). Combating weight-based cyberbullying on Facebook with the dissenter effect. *Cyberpsychology, Behavior, and Social Networking*, *17*(5), 281-286.

Andriopoulos, S. (2011). The sleeper effect: Hypnotism, mind control, terrorism. *Grey Room*, *45*, 88-105.

Arceneaux, K., Gravelle, T. B., Osmundsen, M., Petersen, M. B., Reifler, J., & Scotto, T. J. (2021). Some people just want to watch the world burn: the prevalence, psychology and politics of the 'Need for Chaos'. *Philosophical Transactions of the Royal Society B*, *376*(1822), 20200147.

Bäckström, M., Björklund, F., & Larsson, M. R. (2009). Five-factor inventories have a major general factor related to social desirability which can be reduced by framing items neutrally. *Journal of Research in Personality*, *43*(3), 335-344.

Baer, R. A., & Miller, J. (2002). Underreporting of psychopathology on the MMPI-2: A meta-analytic review. *Psychological Assessment*, *14*(1), 16-37.

Bagby, R. M., & Marshall, M. B. (2003). Positive impression management and its influence on the Revised NEO Personality Inventory: a comparison of analog and differential prevalence group designs. *Psychological Assessment*, *15*(3), 333-342.

Bagby, R. M., Nicholson, R. A., Buis, T., Radovanovic, H., & Fidler, B. J. (1999). Defensive responding on the MMPI-2 in family custody and access evaluations. *Psychological Assessment*, *11*(1), 24-32.

Bargh, J. A., McKenna, K. Y., & Fitzsimons, G. M. (2002). Can you see the real me? Activation and expression of the "true self" on the Internet. *Journal of Social Issues*, *58*(1), 33-48.

Barlińska, J., Szuster, A., & Winiewski, M. (2013). Cyberbullying among adolescent bystanders: Role of the communication medium, form of violence, and empathy. *Journal of Community & Applied Social Psychology*, *23*(1), 37-51.

Ben-Porath, Y. S., & Waller, N. G. (1992). "Normal" personality inventories in clinical assessment: General requirements and the potential for using the NEO Personality Inventory. *Psychological Assessment*, *4*(1), 14-26.

Ben-Ze'ev, A. (2003). Privacy, emotional closeness, and openness in cyberspace. *Computers in Human Behavior*, *19*(4), 451-467.

Bergstrom, K. (2011). “Don’t feed the troll”: Shutting down debate about community expectations on Reddit. com. *First Monday*, *16*(8).

Bianchi, E. C., Hall, E. V., & Lee, S. (2018). Reexamining the Link Between Economic Downturns and Racial Antipathy: Evidence That Prejudice Against Blacks Rises During Recessions. *Psychological Science*, *29*(10), 1584-1597.

Bourdieu, P. (2001). *Masculine domination*. Stanford University Press.

Brigham, J. C. (1993). College students’ racial attitudes. *Journal of Applied Social Psychology*, *23*(23), 1933-1967.

Buchanan, T. "Potential of the Internet for personality research." In *Psychological experiments on the Internet*, pp. 121-140. Academic Press, 2000.

Butz, D. A., & Plant, E. A. (2006). Perceiving outgroup members as unresponsive: implications for approach-related emotions, intentions, and behavior. *Journal of Personality and social Psychology*, *91*(6), 1066.

Calvete, E., Orue, I., Estévez, A., Villardón, L., & Padilla, P. (2010). Cyberbullying in adolescents: Modalities and aggressors’ profile. *Computers in Human Behavior*, *26*(5), 1128-1135.

Carr, C. T., Vitak, J., & McLaughlin, C. (2013). Strength of social cues in online impression formation: Expanding SIDE research. *Communication Research*, *40*(2), 261-281.

Cashel, M. L., Rogers, R., Sewell, K., & Martin-Cannici, C. (1995). The Personality Assessment Inventory (PAI) and the detection of defensiveness. *Assessment*, *2*(4), 333-342.

Cassidy, W., Jackson, M., & Brown, K. N. (2009). Sticks and stones can break my bones, but how can pixels hurt me? Students’ experiences with cyber-bullying. *School Psychology International*, *30*(4), 383-402.

Chau, M., & Xu, J. (2007). Mining communities and their relationships in blogs: A study of online hate groups. *International Journal of Human-Computer Studies*, *65*(1), 57-70.

Christopherson, K. M. (2007). The positive and negative implications of anonymity in Internet social interactions:“On the Internet, Nobody Knows You’re a Dog”. *Computers in Human Behavior*, *23*(6), 3038-3056.

Crandall, C. S., & Eshleman, A. (2003). A justification-suppression model of the expression and experience of prejudice. *Psychological bulletin*, *129*(3), 414.

Crawford, K. (2009). Following you: Disciplines of listening in social media. *Continuum*, *23*(4), 525-535.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*(4), 349-361.

Czopp, A. M., Monteith, M. J., & Mark, A. Y. (2006). Standing up for a change: Reducing bias through interpersonal confrontation. *Journal of Personality and Social Psychology*, *90*(5), 784-793.

Dahlberg, L. (2001). The Internet and democratic discourse: Exploring the prospects of online deliberative forums extending the public sphere. *Information, Communication & Society*, *4*(4), 615-633.

Davis, R. (1999). *The web of politics: The Internet's impact on the American political system*. Oxford University Press.

DeSante, C. D., & Smith, C. W. (2020). Fear, institutionalized racism, and empathy: the underlying dimensions of whites' racial attitudes. *PS: Political Science & Politics*, *53*(4), 639-645.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*(1), 5-18.

DiFonzo, N., & Bordia, P. (2007). Rumor, gossip and urban legends. *Diogenes*, *54*(1), 19-35.

Dilmac, B. (2009). Psychological needs as a predictor of cyber bullying: A preliminary report on college students. *Educational Sciences: Theory and Practice*, *9*(3), 1307-1325.

Donath, J. (1998). Identity and deception in the virtual community. *Communities in Cyberspace* ed. By Marc Smith and Peter Kollock. Routledge.

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology*, *82*(1), 62-70.

Dunton, B. C., & Fazio, R. H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin*, *23*(3), 316-326.

Dyer, R., Green, R., Pitts, M., & Millward, G. (1995). What's the Flaming Problem? or Computer Mediated Communication-Deindividuating or Disinhibiting?. In *BCS HCI*(pp. 289-302).

Fairchild, H. H. (2018). Modern-day racism masks its ugly head. *Social psychology and world peace: A primer*, 213.

Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. *Journal of consumer research*, *20*(2), 303-315.

Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. *The handbook of social psychology*, *2*, 357-411.

Furnham, A. (1990). Faking personality questionnaires: Fabricating different profiles for different purposes. *Current Psychology*, *9*(1), 46-55.

Gaertner, S. L., & Dovidio, J. F. (1977). The subtlety of White racism, arousal, and helping behavior. *Journal of Personality and Social Psychology*, *35*(10), 691-702.

Gaertner, S. L., & Dovidio, J. F. (1981). Racism among the well-intentioned. *Pluralism, racism, and public policy: The search for equality*, 208-222.

Gaertner, S. L., & Dovidio, J. F. (1986). *The aversive form of racism*. San Diego, CA, US: Academic Press.

Görzig, A., & Ólafsson, K. (2013). What makes a bully a cyberbully? Unravelling the characteristics of cyberbullies across twenty-five European countries. *Journal of Children and Media*, *7*(1), 9-27.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464-1476.

Hardaker, C. (2010). Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *Journal of Politeness Research, 6(2),* 215-242.

Harman, J. P., Hansen, C. E., Cochran, M. E., & Lindsey, C. R. (2005). Liar, liar: Internet faking but not frequency of use affects social skills, self-esteem, social anxiety, and aggression. *CyberPsychology & Behavior*, *8*(1), 1-6.

Hawdon, J., Oksanen, A., & Räsänen, P. (2017). Exposure to online hate in four nations: A cross-national consideration. *Deviant Behavior*, *38*(3), 254-266.

Hinduja, S., & Patchin, J. W. (2008). Cyberbullying: An exploratory analysis of factors related to offending and victimization. *Deviant Behavior*, *29*(2), 129-156.

Huang, Y. Y., & Chou, C. (2010). An analysis of multiple factors of cyberbullying among junior high school students in Taiwan. *Computers in Human Behavior*, *26*(6), 1581-1590.

Jardina, A., & Piston, S. (2019). Racial prejudice, racial identity, and attitudes in political decision making. In *Oxford Research Encyclopedia of Politics*.

Johnson-Agbakwu, C. E., Ali, N. S., Oxford, C. M., Wingo, S., Manin, E., & Coonrod, D. V. (2020). Racism, COVID-19, and Health Inequity in the USA: a Call to Action. *Journal of racial and ethnic health disparities*, 1-7.

Joinson, A. (1999). Social desirability, anonymity, and Internet-based questionnaires. *Behavior Research Methods, Instruments, & Computers*, *31*(3), 433-438.

Joinson, A. N. (2001). Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity. *European Journal of Social Psychology*, *31*(2), 177-192.

Joinson, A. N. (2007). Disinhibition and the Internet. In *Psychology and the Internet (Second Edition)* (pp. 75-92). Academic Press. Alberta.

Jones-Jang, S. M., Mortensen, T., & Liu, J. (2019). Does media literacy help identification of fake news? Information literacy helps, but other literacies don't. *American Behavioral Scientist*, *65*(2), 371-388.

Keum, B. T., & Miller, M. J. (2018). Racism on the Internet: Conceptualization and recommendations for research. *Psychology of violence*, *8*(6), 782.

Kiesler, S., Siegel, J., & McGuire, T. W. (1984). Social psychological aspects of computer-mediated communication. *American Psychologist*, *39*(10), 1123-1130.

Kowalski, R. M., & Limber, S. P. (2007). Electronic bullying among middle school students. *Journal of Adolescent Health*, *41*(6), S22-S30.

Lapidot-Lefler, N., & Barak, A. (2012). Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior*, *28*(2), 434-443.

Laustsen, L., & Petersen, M. B. (2015). Does a competent leader make a good friend? Conflict, ideology and the psychologies of friendship and followership. *Evolution and Human Behavior*, *36*(4), 286-293.

Lawless, T. J. & Saucier, D. A. (submitted a). Power through anonymity? Perceptions of prejudice on social media. *Computers in Human Behavior*.

Lawless, T. J. & Saucier, D. A. (submitted b). The wild wild web: Anonymity and racial prejudice in online culture. *Computers in Human Behavior*.

Lawless, T. J. & Saucier, D. A. (in preparation). Taking a SIDE: Effects of deindividuation and group feedback on perceptions of prejudice on social media.

Lenhart, A., Madden, M., Smith, A., Purcell, K., Zickuhr, K., & Rainie, L. (2011). Teens, Kindness and Cruelty on Social Network Sites: How American Teens Navigate the New World of" Digital Citizenship". *Pew Internet & American Life Project*.

Li, T. B. Q. (2005). Cyber-harassment: A study of a new method for an old behavior. *Journal of Educational Computing Research*, *32*(3), 265-277.

Linder, C. (2015). Navigating guilt, shame, and fear of appearing racist: A conceptual model of antiracist white feminist identity development. *Journal of College Student Development*, *56*(6), 535-550.

Locey, M. L., & Rachlin, H. (2015). Altruism and anonymity: A behavioral analysis. *Behavioural Processes*, *118*, 71-75.

Mandalaywala, T. M., Amodio, D. M., & Rhodes, M. (2018). Essentialism promotes racial prejudice by increasing endorsement of social hierarchies. *Social Psychological and Personality Science*, *9*(4), 461-469.

Martin, J. K., Pescosolido, B. A., & Tuch, S. A. (2000). Of fear and loathing: The role of'disturbing behavior,'labels, and causal attributions in shaping public attitudes toward people with mental illness. *Journal of Health and Social Behavior*, 208-223.

McCright, A. M., & Dunlap, R. E. (2017). Combatting misinformation requires recognizing its types and the factors that facilitate its spread and resonance. *Journal of Applied Research in Memory and Cognition*, *6*(4), 389-396.

McManus, J. L., Saucier, D. A., O'Dea, C. J., & Bernard, D. L. (2019). Aversive Affect Versus Racism as Predictors of Racial Discrimination in Helping. *Basic and Applied Social Psychology*, *41*(4), 230-253.

Miller, S. S., O'Dea, C. J., Lawless, T. J., & Saucier, D. A. (2019). Savage or satire: Individual differences in perceptions of disparaging and subversive racial humor. *Personality and Individual Differences*, *142*, 28-41.

Miller, S. S., O'Dea, C. J., & Saucier, D. A. (2021). "I can't breathe": Lay conceptualizations of racism predict support for Black Lives Matter. *Personality and individual differences*, *173*, 110625.

Miller, S. S., Peacock, N. K., & Saucier, D. A. (2021). Propensities to make attributions to prejudice and (mis) perceptions of racism in the context of police shootings. *Personality and individual differences*, *182*, 111088.

Miller, S. S., & Saucier, D. A. (2018). Individual differences in the propensity to make attributions to prejudice. *Group Processes & Intergroup Relations*, *21*(2), 280-301.

Miller, S. S., Martens, A. L., & Saucier, D. A. (2017). Attributions to Prejudice: Collective Anger and Action. *Understanding Angry Groups: Multidisciplinary Perspectives on Their Motivations and Effects on Society*, 29-40.

Milner, A., & Franz, B. (2020). Anti-black attitudes are a threat to health equity in the United States. *Journal of Racial and Ethnic Health Disparities*, *7*(1), 169-176.

Moor, P. J., Heuvelman, A., & Verleur, R. (2010). Flaming on youtube. *Computers in Human Behavior*, *26*(6), 1536-1546.

Moore, M. J., Nakano, T., Enomoto, A., & Suda, T. (2012). Anonymity and roles associated with aggressive posts in an online forum. *Computers in Human Behavior*, *28*(3), 861-867.

Moral-Toranzo, F., Canto-Ortiz, J., & Gómez-Jacinto, L. (2007). Anonymity effects in computer-mediated communication in the case of minority influence. *Computers in Human Behavior*, *23*(3), 1660-1674.

Ng, E. (2020). No grand pronouncements here...: Reflections on cancel culture and digital media participation. *Television & New Media*, *21*(6), 621-627.

Nichols, D. S., & Greene, R. L. (1997). Dimensions of deception in personality assessment: The example of the MMPI-2. *Journal of Personality Assessment*, *68*(2), 251-266.

Nogami, T., & Takai, J. (2008). Effects of anonymity on antisocial behavior committed by individuals. *Psychological Reports*, *102*(1), 119-130.

O'Dea, C. J., Bueno, A. M. C., & Saucier, D. A. (2018). Social vigilantism and the extremity, superiority, and defense of attitudes toward climate change. *Personality and Individual Differences*, *130*, 83-91.

O'Dea, C. J., Zhu, Q., & Saucier, D. A. (in preparation). Engaged and upset: Social vigilantism, positive and negative affect, and resistance to attitude challenges.

Pasveer, K. A., & Ellard, J. H. (1998). The making of a personality inventory: Help from the WWW. *Behavior Research Methods, Instruments, & Computers*, *30*(2), 309-313.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of personality and social psychology*, *46*(3), 598.

Pauls, C. A., & Crost, N. W. (2004). Effects of faking on self-deception and impression management scales. *Personality and Individual Differences*, *37*(6), 1137-1151.

Pauls, C. A., & Crost, N. W. (2005). Cognitive ability and self-reported efficacy of self-presentation predict faking on personality measures. *Journal of Individual Differences*, *26*(4), 194-206.

Payne, B. K., & Hannay, J. W. (2021). Implicit bias reflects systemic racism. *Trends in cognitive sciences*, *25*(11), 927-936.

Petersen, M. B., Osmundsen, M., & Arceneaux, K. (2018). A "need for chaos" and the sharing of hostile political rumors in advanced democracies. *American Political Science Association, Boston, MA*, *30*.

Petrocelli, J. V., Seta, C. E., & Seta, J. J. (2023). Lies and bullshit: The negative effects of misinformation grow stronger over time. *Applied Cognitive Psychology*, *37*(2), 409-418.

Plant, E. A. (2004). Responses to interracial interactions over time. *Personality and Social Psychology Bulletin*, *30*(11), 1458-1471.

Plant, E. A., & Butz, D. A. (2006). The causes and consequences of an avoidance-focus for interracial interactions. *Personality and Social Psychology Bulletin*, *32*(6), 833-846.

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology*, *75*(3), 811.

Postmes, T., & Baym, N. (2005). Intergroup dimensions of the Internet. *Intergroup communication: Multiple Perspectives*, *2*, 213-240.

Postmes, T., & Spears, R. (2002). Behavior online: Does anonymous computer communication reduce gender inequality?. *Personality and Social Psychology Bulletin*, *28*(8), 1073-1083.

Postmes, T., Spears, R., & Lea, M. (1998). Breaching or building social boundaries? SIDE-effects of computer-mediated communication. *Communication Research*, *25*(6), 689-715.

Postmes, T., Spears, R., Sakhel, K., & De Groot, D. (2001). Social influence in computer-mediated communication: The effects of anonymity on group behavior. *Personality and Social Psychology Bulletin*, *27*(10), 1243-1254.

Price, M., & Dalgleish, J. (2010). Cyberbullying: Experiences, impacts and coping strategies as described by Australian young people. *Youth Studies Australia*, *29*(2), 51-63.

Raimi, K. T., & Leary, M. R. (2014). Belief superiority in the environmental domain: Attitude extremity and reactions to fracking. *Journal of Environmental Psychology*, *40*, 76-85.

Räsänen, P., Hawdon, J., Holkeri, E., Keipi, T., Näsi, M., & Oksanen, A. (2016). Targets of online hate: Examining determinants of victimization among young Finnish Facebook users. *Violence and Victims*, *31*(4), 708-718.

Reader, B. (2012). Free press vs. free speech? The rhetoric of "civility" in regard to anonymous online comments. *Journalism & Mass Communication Quarterly*, *89*(3), 495-513.

Reeckman, B., & Cannard, L. (2009). Cyberbullying: a TAFE perspective. *Youth Studies Australia*, *28*(2), 41-54.

Retzlaff, P., Sheehan, E., & Fiel, A. (1991). MCMI-II report style and bias: Profile and validity scales analyses. *Journal of Personality Assessment*, *56*(3), 466-477.

Romano, A. (2019). Why we can't stop fighting about cancel culture. *Vox Magazine. URL https://www.vox.com/culture/2019/12/30/20879720/what-is-cancel-culture-explained-history-debate*.

Runions, K. C., (2015). Online moral disengagement, cyberbullying, and cyber-aggression. *Cyberpsychology, Behavior, and Social Networking*, *18*(7), 400-405.

Saucier, D. A., Hockett, J. M., O'Dea, C. J., & Miller, S. S. (2016). The racism justification hypothesis and attitudes toward hate crime legislation. *The Psychology of Hate Crimes as Domestic Terrorism: US and Global Issues [3 volumes]: US and Global Issues*, 283.

Saucier, D. A., Hockett, J. M., & Wallenberg, A. S. (2008). The impact of racial slurs and racism on the perceptions and punishment of violent crime. *Journal of interpersonal violence*, *23*(5), 685-701.

Saucier, D. A., Hockett, J. M., Zanotti, D. C., & Heffel, S. (2010). Effects of racism on perceptions and punishment of intra-and interracial crimes. *Journal of Interpersonal Violence*, *25*(10), 1767-1784.

Saucier, D. A., & Miller, C. T. (2003). The persuasiveness of racial arguments as a subtle measure of racism. *Personality and Social Psychology Bulletin*, *29*(10), 1303-1315.

Saucier, D. A., Miller, C. T., & Doucet, N. (2005). Differences in helping whites and blacks: A meta-analysis. *Personality and Social Psychology Review*, *9*(1), 2-16.

Saucier, D. A., Miller, S. S., Martens, A. L., & O'Dea, C. J. (2017). Overt racism. In A. M. Czopp & A. W. Blume (Eds.), *Social Issues in Living Color: Challenges and Solutions from the Perspective of Ethnic Minority Psychology: Societal and Global Issues*, 77-102.

Saucier, D. A., Smith, S. J., & Lawless, T. J. (2021). Ardent, but informed? Social vigilantism and the dissemination and defense of political decisions. *Personality and Individual Differences*, *179*, 110887.

Saucier, D. A., & Webster, R. J. (2010). Social vigilantism: Measuring individual differences in belief superiority and resistance to persuasion. *Personality and Social Psychology Bulletin*, *36*(1), 19-32.

Saucier, D. A., Webster, R. J., Hoffman, B. H., & Strain, M. L. (2014). Social vigilantism and reported use of strategies to resist persuasion. *Personality and Individual Differences*, *70*, 120-125.

Scandell, D. J., & Wlazelek, B. G. (1996). Self-presentation strategies on the NEO-Five Factor Inventory: Implications for detecting faking. *Psychological Reports*, *79*(3_suppl), 1115-1121.

Schnurr, M. P., Mahatmya, D., & Basche III, R. A. (2013). The role of dominance, cyber aggression perpetration, and gender on emerging adults' perpetration of intimate partner violence. *Psychology of violence*, *3*(1), 70.

Scott, C. R. (1998). The impact of physical and discursive anonymity on group members' multiple identifications during computer-supported decision making. *Western Journal of Communication (includes Communication Reports)*, *63*(4), 456-487.

Shelton, J. N. (2003). Interpersonal concerns in social encounters between majority and minority group members. *Group Processes & Intergroup Relations*, *6*(2), 171-185.

Slonje, R., Smith, P. K., & Frisén, A. (2012). Processes of cyberbullying, and feelings of remorse by bullies: A pilot study. *European Journal of Developmental Psychology*, *9*(2), 244-259.

Smith, M. A., & Leigh, B. (1997). Virtual subjects: Using the Internet as an alternative source of subjects and research environment. *Behavior Research Methods, Instruments, & Computers*, *29*(4), 496-505.

Son Hing, L. S., Chung-Yan, G. A., Grunfeld, R., Robichaud, L. K., & Zanna, M. P. (2005). Exploring the discrepancy between implicit and explicit prejudice: A test of aversive racism theory. In *Biennial meeting of the Society for the Psychological Study of Social Issues., Jun, 2002, Toronto, ON, Canada. March 2003, Sydney, Australia*. Cambridge University Press.

Sparrow, J. A. (2019). This Is America: Examining the Influence of White Privilege and the Propensity to Make Attributions to Prejudice on the Acceptability of Racial Microaggressions.

Spears, R., & Lea, M. (1992). *Social influence and the influence of the 'social 'in computer-mediated communication*. Harvester Wheatsheaf.

Sproull, L., & Kiesler, S. (1986). Reducing social context cues: Electronic mail in organizational communication. *Management science*, *32*(11), 1492-1512.

Sticca, F., & Perren, S. (2012). Is cyberbullying worse than traditional bullying? Examining the differential roles of medium, publicity, and anonymity for the perceived severity of bullying. *Journal of Youth and Adolescence*, *42*(5), 739-750.

Stratmoen, E., Lawless, T. J., & Saucier, D. A. (2019). Taking a knee: Perceptions of NFL player protests during the National Anthem. *Personality and Individual Differences*, *137*, 204-213.

Sue, D. W. (2013). Race talk: The psychology of racial dialogues. *American Psychologist*, *68*(8), 663.

Sugarman, D. B., & Willoughby, T. (2013). Technology and violence: Conceptual issues raised by the rapidly changing social environment. *Psychology of Violence*, *3*(1), 1.

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & behavior*, *7*(3), 321-326.

Tidwell, L. C., & Walther, J. B. (2002). Computer-mediated communication effects on disclosure, impressions, and interpersonal evaluations: Getting to know one another a bit at a time. *Human Communication Research*, *28*(3), 317-348.

Udris, R. (2014). Cyberbullying among high school students in Japan: Development and validation of the Online Disinhibition Scale. *Computers in Human Behavior*, *41*, 253-261.

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement*, *59*(2), 197-210.

Voggeser, B. J., Singh, R. K., & Göritz, A. S. (2018). Self-control in online discussions: Disinhibited online behavior as a failure to recognize social cues. *Frontiers in psychology*, 2372.

Walther, J. B., & Bazarova, N. N. (2007). Misattribution in virtual groups: The effects of member distribution on self-serving bias and partner blame. *Human Communication Research*, *33*(1), 1-26.

Waltman, M. S. (2018). The normalizing of hate speech and how communication educators should respond. *Communication Education, 67(2)*, 259-265.

Wang, G., Wang, B., Wang, T., Nika, A., Zheng, H., & Zhao, B. Y. (2014, November). Whispers in the dark: analysis of an anonymous social network. In *Proceedings of the 2014 Conference on Internet Measurement Conference* (pp. 137-150). ACM.

Willard, N. E. (2007). *Cyberbullying and cyberthreats: Responding to the challenge of online social aggression, threats, and distress*. Research Press.

Williams, D., & Skoric, M. (2005). Internet fantasy violence: A test of aggression in an online game. *Communication Monographs*, *72*(2), 217-233.

Wright, M. F. (2013). The relationship between young adults' beliefs about anonymity and subsequent cyber aggression. *Cyberpsychology, Behavior, and Social Networking*, *16*(12), 858-862.

Wright, M. F., Harper, B. D., & Wachs, S. (2018). The associations between cyberbullying and callous-unemotional traits among adolescents: The moderating effect of online disinhibition. *Personality and Individual Differences*.

Yadon, N., & Piston, S. (2019). Examining whites' anti-black attitudes after Obama's presidency. *Politics, Groups, and Identities, 7*(4), 794-814.

Appendix A - Demographics Materials (coded values in parentheses)

Age

_____

What is your ethnicity

- ○ White (1)
- ○ Black (2)
- ○ Hispanic/Latino/a/x (3)
- ○ Asian (4)
- ○ Native American (5)
- ○ Pacific Islander (6)
- ○ Multiracial (7)
- ○ I prefer to describe my ethnicity using my own language: (8) _____
- ○ I prefer not to disclose (9)

What is your current gender identity? (Please select all that apply)

- [ ] Female (1)
- [ ] Male (2)
- [ ] Genderqueer (3)
- [ ] Genderfluid (4)
- [ ] Transgender (5)
- [ ] Transgender Feminine (6)
- [ ] Transgender Masculine (7)
- [ ] Agender (8)
- [ ] Androgynous (9)
- [ ] Two Spirit (10)
- [ ] Demigender (11)
- [ ] Questioning or unsure (12)
- [ ] I prefer to describe my gender using my own language: (13) _____
- [ ] I prefer not to disclose. (14)

Although it is often difficult to summarize one's political, economic, social, and religious views in a single word or phrase, please indicate your overall political viewpoint on the scale below.

- ○ 1 = Very Liberal
- ○ 2
- ○ 3
- ○ 4
- ○ 5
- ○ 6
- ○ 7
- ○ 8
- ○ 9 = Very Conservative

Appendix B - Stimuli

Original Posts (prejudiced)

1. Passed four Black guys on my way home and I've never been so scared in my life.
2. Not sure how anyone can stand to be around Black people. They're so loud.
3. Welfare … so all the Black people can continue to be lazy and profit.
4. Wonder if there's anything Black people are good at besides sports.
5. Black people whine and complain about being "oppressed" yet sit at home and collect welfare. It's called hard work!
6. My assigned partner in my Chem lab is this Black guy, so I guess I'm screwed.
7. Black people's logic: If at first you don't succeed, play the victim.
8. Black people rioting = Real life Planet of the Apes

Confrontational Comments:

1. You're a joke if you actually think this.
2. Dude, this is racist you bigot!
3. You're the worst type of person.
4. Do you understand how racist this is?

Appendix C - Criterion Measures

Note: All measures rated on 1 (*Strongly Disagree*) to 9 (*Strongly Agree*) Likert-type scales unless otherwise designated.

Included with each post:

This post is racist.

This post is meant to harm.

This post is meant to be hurtful.

This post is intended to be cruel.

The person who created this post is racist.

The person who created this post is cruel.

The person who created this post is hurtful.

The person who created this post meant to harm.

This post is meant to be taken honestly.

This post reflects the true beliefs of the person who made it.

The person who created this post believes that this post reflects reality.

The person who created this post wants others to believe that this post reflects reality.

The person who created this post just wants attention.

For use only in posts that included comments:

The commenter who confronted this post did the right thing.

The commenter shouldn't have said anything

For use only in Study 2:

I would leave a positive comment on this post.

I would leave a negative comment on this post.

I would block the user who made the original post.

I would report the user who made the original post.

If you would comment on this post, what would you say? (open-ended item)

_____

For use only on Facebook mock posts:

*I would "like" this post.

**I would angry face react this post.

For use only on Reddit mock posts:

*I would upvote this post.

**I would downvote this post.

*For analysis purposes, these items will be combined into “positive reaction”

**For analysis purposes, these items will be combined into “negative reaction”

Appendix D - Individual Difference Measures

Note: All measures rated on 1 (*Strongly Disagree*) to 9 (*Strongly Agree*) Likert-type scales unless otherwise designated.

Propensity to Make Attributions to Prejudice Scale (Miller & Saucier, 2016; including two additional attention check items)

1. People discriminate against people who are not like them.
2. Racist behavior is more widespread than people think it is.
3. Other people treat minorities based on stereotypes.
4. You'll see lots of racism if you look for it.
5. Racial minorities are too worried about being discriminated against.
6. Racial minorities are too sensitive about stereotypes.
7. Minorities today are overly worried about being victims of racism.
8. People are overly concerned about racial issues.
9. I think about why racial minorities are treated stereotypically.
10. Please mark 8 for this answer
11. I think about whether people act in a prejudiced or discriminatory manner.
12. Please select eight
13. I consider whether people's actions are prejudiced or discriminatory.
14. I am on the lookout for instances of prejudice or discrimination.
15. I am quick to recognize prejudice.
16. My friends think I'm good at spotting racism.
17. I find that prejudice and discrimination are pretty easy to spot.

Attitudes Toward Blacks Scale (Brigham, 1993)

1. I enjoy a funny racial joke, even if some people might find it offensive.
2. If I had a chance to introduce Black visitors to my friends and neighbors, I would be pleased to do so.
3. I would rather not have Blacks live in the same apartment building I live in.
4. Racial integration (of schools, businesses, residences, etc.) has benefited both Whites and Blacks.
5. I would probably feel somewhat self-conscious dancing with a Black in a public place.
6. I think that Black people look more similar to each other than White people do.
7. It would not bother me if my new roommate was Black.
8. Inter-racial marriage should be discouraged to avoid the "who-am-I?" confusion which the children feel.
9. If a Black were put in charge of me, I would not mind taking advice and direction from him or her.
10. Generally, Blacks are not as smart as Whites.
11. The federal government should take decisive steps to override the injustices Blacks suffer at the hands of local authorities.
12. It is likely that Blacks will bring violence to neighborhoods when they move in.
13. Black and White people are inherently equal.
14. I get very upset when I hear a White make a prejudicial remark about Blacks.
15. I worry that in the next few years I may be denied my application for a job or promotion because of the preferential treatment given to minority group members.
16. I favor open housing laws that allow more racial integration of neighborhoods.

17. Black people are demanding too much too fast in their push for equal rights.
18. I would not mind at all if a Black family with about the same income and education as me moved in next door.
19. Whites should support Blacks in their struggle against discrimination and segregation.
20. Some Blacks are so touchy about race that it is difficult to get along with them.

Social Desirability Scale (Crowne & Marlowe, 1960; measured as True/False)

1. Before voting, I thoroughly investigate the qualifications of all the candidates.
2. I never hesitate to go out of my way to help someone in trouble.
3. It is sometimes hard for me to go on with my work if I am not encouraged.
4. I have never intensely disliked anyone.
5. On occasion I have had doubts about my ability to succeed in life.
6. I sometimes feel resentful when I don't get my way.
7. I am always careful about my manner of dress.
8. My table manners at home are as good as when I eat out in a restaurant.
9. If I could get into a movie without paying and be sure I was not seen I would probably do it.
10. On a few occasions, I have given up doing something because I thought too little of my ability.
11. I like to gossip at times.
12. There have been times when I felt like rebelling against people in authority even though I knew they were right.
13. No matter who I'm talking to, I'm always a good listener.
14. I can remember "playing sick" to get out of something.
15. There have been occasions when I took advantage of someone.
16. I'm always willing to admit it when I make a mistake.
17. I always try to practice what I preach.
18. I don't find it particularly difficult to get along with loud mouthed, obnoxious people.
19. I sometimes try to get even rather than forgive and forget.
20. When I don't know something I don't at all mind admitting it.
21. I am always courteous, even to people who are disagreeable.
22. At times I have really insisted on having things my own way.
23. There have been occasions when I felt like smashing things.
24. I would never think of letting someone else be punished for my wrong-doings.
25. I never resent being asked to return a favor.
26. I have never been irked when people expressed ideas very different from my own.
27. I never make a long trip without checking the safety of my car.

28. There have been times when I was quite jealous of the good fortune of others.
29. I have almost never felt the urge to tell someone off.
30. I am sometimes irritated by people who ask favors of me.
31. I have never felt that I was punished without cause.
32. I sometimes think when people have a misfortune they only got what they deserved.
33. I have never deliberately said something that hurt someone's feelings.

Need for Chaos Scale (Petersen et al., 2018)

1. I get a kick when natural disasters strike in foreign countries.
2. I fantasize about a natural disaster wiping out most of humanity such that a small group of people can start all over.
3. I think society should be burned to the ground.
4. When I think about our political and social institutions, I cannot help thinking "just let them all burn."
5. We cannot fix the problems in our social institutions, we need to tear them down and start over.
6. I need chaos around me - it is too boring if nothing is going on.
7. Sometimes I just feel like destroying beautiful things.
8. There is no right and wrong in the world.

Social Vigilantism Scale (Saucier & Webster, 2010; including one additional attention check item)

1. I feel as if it is my duty to enlighten other people.
2. I feel that my ideas should be used to educate others.
3. I feel a social obligation to voice my opinion.
4. I need to win any argument about how people should live their lives.
5. The people who are more intelligent and informed have a responsibility to educate the people around them who are less intelligent and informed.
6. I like to imagine myself in a position of authority so that I could make the important decisions around here.
7. Select 3
8. I try to get people to listen to me, because what I have to say makes a lot of sense.
9. Some people just believe stupid things.
10. There are a lot of ignorant people in society.
11. I think that some people need to be told that their point of view is wrong.
12. If everyone saw things the way that I do, the world would be a better place.
13. It frustrates me that many people fail to consider the finer points of an issue when they take a side.
14. I often feel that other people do not base their opinions on good evidence.
15. I frequently consider writing a “letter to the editor.”