2020

Antecedents of support for social media content moderation and platform regulation: the role of presumed effects on self and others

Martin J. Riedl

Kelsey Whipple

Ryan Wallace

The Version of Record of this manuscript has been published and is available in
Information, Communication & Society (2021), https://doi.org/10.1080/1369118X.2021.1874040

Antecedents of support for social media content moderation and platform regulation: the role of presumed effects on self and others

Martin J. Riedl, Kelsey N. Whipple, and Ryan Wallace

Abstract

This study examines support for regulation of and by platforms and provides insights into public perceptions of platform governance. While much of the public discourse surrounding platforms evolves at a policy level between think tanks, journalists, academics and political actors, little attention is paid to how people think about regulation of and by platforms. Through a representative survey study of U.S. internet users (N = 1,022), we explore antecedents of support for social media content moderation by platforms, as well as for regulation of social media platforms by the government. We connect these findings to presumed effects on self (PME1) and others (PME3), concepts that lie at the core of third-person effect (TPE) and influence of presumed influence (IPI) scholarship. We identify third-person perceptions for social media content: Perceived negative effects are stronger for others than for oneself. A first-person perception operates on the platform level: The beneficial effects of social media platforms are perceived to be stronger for the self than for society. At the behavioral level, we identify age, education, opposition to censorship, and perceived negative effects of social media content on others (PME3) as significant predictors of support for content moderation. Concerning support for regulation of platforms by the government, we find significant effects of opposition to censorship, perceived intentional censorship, frequency of social media use, and trust in platforms. We argue that stakeholders involved in platform governance must take more seriously the attitudes of their constituents.

Keywords: content moderation, social media, platform regulation, third-person effect, survey, free speech

Introduction

Content moderation and platform regulation are having a moment. While just a few years ago the profession of content moderators was largely unknown to the public, profiles in prominent news outlets (e.g., A. Chen, 2017) have raised awareness for this critical digital labor issue. At the same time, congressional hearings on Section 230 of the Communications Decency Act, a U.S. law regulating the liability of internet platforms (Medeiros, 2017), as well as hearings in which Facebook CEO Mark Zuckerberg provided testimony on Capitol Hill, have thrust social media platforms to the front and center of ongoing political debate.

The spotlight on platform regulation, perceived biases of social media platforms (Allen & VandeHei, 2019), and the ways in which platforms are enmeshed in major social events including elections (e.g., Allcott & Gentzkow, 2017) have exposed the complicated relationships between platforms and society. Today, people question how platforms govern online content production and distribution. The Pew Research Center reports that about 70% of Americans think social media platforms likely censor political viewpoints (Smith, 2018). Content takedowns, however, are foremost a risk-mitigation project: Certain types of content, such as

The Version of Record of this manuscript has been published and is available in
Information, Communication & Society (2021), https://doi.org/10.1080/1369118X.2021.1874040

terrorist propaganda, child pornography, or hate speech are more harmful to society than others. Lawmakers and activists alike are calling for platforms to remove such harmful content with increased speed and transparency.

Ongoing discussions about platform governance (Gorwa, 2019) are valuable; however, they often ignore critical questions about what users and nonusers think. This study explores antecedents of support for regulation by and of platforms in light of perceived effects of social media in society from an audience perspective.

Through a representative, cross-sectional survey study of the United States internet population (N = 1,022), we explore these issues in two critical ways: First, we investigate and survey factors impacting support for content moderation by platforms and presumed effects of social media content on the self vis-à-vis other people. Second, we explore factors impacting support for increased government regulation for platforms, and presumed effects related to the impact of platforms on society vis-à-vis the self. In doing so, we explore the perceptual claims of the third-person effect (TPE). Our inquiry is informed by scholarship on TPE and the influence of presumed influence (IPI) framework, as well as more recent scholarship emphasizing the importance of considering the impact of perceived effects on self (PME1) vis-à-vis perceived effects on others (PME3) as independent variables.

Literature Review

Content moderation as regulation by platforms

Content moderation relates to "*governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse* [italics in original]" (Grimmelmann, 2015, p. 47). Moderation is contingent on user-generated content, or UGC, and often outsourced and opaque by design (Roberts, 2019). UGC can be moderated ex-ante, or ex-post (Klonick, 2017), the latter of which can be done by moderation teams wading through forums, or by responding to audience members flagging content (Naab et al., 2018). Moderation is further enabled by computational means, for example through machine learning and filtering (Myers West, 2018). Moderation can also be a volunteer activity; Matias (2019) describes content moderation that is carried out by volunteers as a type of *civic labor*. Content moderation is one of the most pervasive functions of social media platforms, as it is instrumental to shaping the content regimes that users encounter. Gillespie (2018) describes moderation as the main commodity platforms offer. Content moderation also confronts societies with important questions about civic liberties, such to what extent citizens want private actors like social media platforms to police content, and how freedom of expression factors into content governance regimes. In the United States, Section 230 of the Communications Decency Act governs the liability of platforms (Medeiros, 2017) and makes it possible for platforms to voluntarily conduct moderation. When content moderation is framed as regulation, law scholars such as Klonick (2017) refer to platforms with terms such as the *new governors*. The underlying assumption of moderation is that some content is considered to be detrimental to users. To explore attitudes and perceptions surrounding social media content and moderation, we turn our attention to perceptions of such effects on self (PME1) as well as perceived effects on others (PME3), and

two frameworks associated with these perceptions: The third-person effect (TPE) and the influence of presumed influence (IPI) model.

Third-person effect, presumed influence on self and others, and social media

The third-person effect (TPE) of communication arises when an individual exposed to a mass media message perceives that message as being more impactful or persuasive to others than to him or herself (Davison, 1983). The third-person effect also predicts that people make decisions based on their assumptions of media impact (Gunther, 1991). For example, they might support restricting or censoring media messages (Salwen & Dupagne, 1999) to protect others who might be influenced. Researchers have examined the third-person effect in the context of key domains of online content and content moderation, including internet pornography (Lee & Tamborini, 2005; Lo & Wei, 2002), Facebook (Paradise & Sullivan, 2012; Tsay-Vogel, 2016), and online comments (Chen & Ng, 2016). More recently, researchers have extended the concept to the domain of social media platforms. For example, Betts et al. (2019) found that people perceive others as more likely to experience cyber bullying than themselves. In the domain of *fake news*, that is, news propagating false information, stronger third-person perceptions have been shown to lead to lower support for regulation (Jang & Kim, 2018). Tsay-Vogel (2016) tested a perceived *Facebook effect* from a third-person perspective and confirmed that people perceive Facebook as affecting others more than themselves. However, her measurement conflates content effects and platform effects, issues we are keen to distinguish in this present study.

The literature on TPE propagates two core hypotheses: The first one assumes that there is a perceptual difference between how a person thinks they might be affected by media content versus how others might be. This perceptual difference is widely established and so is its reverse, called a first-person effect, which appears, "[w]hen a presumed media effect is socially desirable" (Baek et al., 2019, p. 303). While a large majority of research focused on perceptions of undesirable content, first-person perceptions are important to consider when contemplating socially desirable content, or, as this study proposes, infrastructures such as social media platforms. The second hypothesis of TPE posits that the difference between perceived effects on others (PME3) and on self (PME1) impacts support for regulation – a behavioral claim that has been contested (Chung & Moon, 2016). In line with the established perceptual hypothesis, we posit:

> **H1.** Participants will perceive the negative effects of social media content to be stronger for others than for themselves.

The third-person effect has long been examined alongside censorship (Davison, 1983). Using it, censors and those who support censorship of some type of media content or another can claim that they are acting to protect the interests of third parties who are more susceptible to media messages than they are (Davison, 1983; Salwen & Dupagne, 1999). Concerning the behavioral hypothesis, researchers have found inconclusive connections between third-person effects and support for regulation of Facebook (Paradise & Sullivan, 2012), whereas other research related to sexual content has found strong connections between third-person effects and support for censorship (Chia et al., 2006).

More recently, researchers including Jang and Kim (2018) have not found a link between third-person effects and support for government regulation. Cheng and Chen (2020) suggest that this lack of relationship may be based on social media users' own wariness of government regulation. In this case, examining the influence of presumed influence (IPI; Gunther & Storey, 2003) on others may be more useful in understanding attitudes toward regulation. Indeed, looking more closely at users' presumptions of the negative effect of media on others — rather than on the difference between self and others — shows a clearer link to support of regulation (Cheng & Chen, 2020; Gunther, 1991). Baek et al. (2019) found that people were more likely to support regulation as an outcome when they believed that fake news had an influence on both other people *and* themselves. Chung and Moon (2016) found that the perception that others are influenced (PME3) is a stronger predictor for support of censorship than the perception that oneself is influenced (PME1).

Another strand of research, coined the *diamond* method (Schmierbach et al., 2011, Neuwirth & Frederick, 2002) proposes exploring a second-person effect (SPE), which relates to a summative term of perceived effects on self and others (PME1 + PME3) (Neuwirth & Frederick, 2002; Sun et al., 2008). This is of interest because such a formula "controls for the effects of general perceptions of strong media effects" (Schmierbach et al., 2011, p. 311), and because second-person perceptions have been found to be strong predictors of behavior (Neuwirth & Frederick, 2002).

When considering explanations for these effects, proposed underlying psychological mechanisms generally fall into one of two categories: those that are cognition-based and those that are motivation-based (Nan, 2007). Cognitive explanations such as Gunther and Mundy's (1993) biased optimism build on the human tendency to perceive the world through a dichotomous and often self-serving lens—one that separates the self and others within the mind. Alternatively, popular motivational explanations (Meirick, 2005) suggest that individuals self-enhance to perceive themselves as less vulnerable to media effects.

We root our study in a conceptualization that assumes social media use entails serendipitous and unavoidable exposure to what Jack (2019) refers to as *wicked content,* which is "recognizably problematic, even if said content's veracity, its provenance, and the intent with which it was distributed, are uncertain" (p. 436). Just how problematic content may be lies in the eye of the beholder, though possibilities to imagine what could constitute *bad, wicked,* or otherwise *harmful* content are endless and leave ample room to platforms for intervention and moderation. Research documents the powerful role that hate speech (e.g., Pohjonen, 2019), online incivility (e.g., G. M. Chen, 2017) and disinformation or propaganda (e.g., Woolley & Howard, 2018) can play in online spaces, rendering considerations of harm on both personal and societal levels. Exposure to incivility, for instance, can lead to emotional exhaustion (Riedl et al., 2020), and emotional distress (Lee-Won et al., 2019). Incivility can also lead to anger when someone's in-group is targeted (Gervais, 2015).

Research indicates that 45% of Americans think that technology companies have a "responsibility to protect the public from objectionable content" (Ballard, 2019, n. p.). At the same time, 41% of Americans think that "[r]removing user-created content from social media sites is suppressing free speech" (Ballard, 2019, n. p.).

Given how social media platforms are implicated in interference in democratic elections in the U.S. and abroad and the strategic spread of misinformation internationally, the responsibility of tech companies stretches well beyond protecting the public from objectionable content (e.g., Allcott & Gentzkow, 2017). Social media platforms, through their user agreements and content moderation, largely govern what users can do and say on their platforms. However, while significant interest has recently concentrated on structural questions of governance, how platforms shape discourse, and what content should be freely shared, less attention is paid to user perceptions. Because the salience of particular content on social media may not only shape what audiences think but can also actively shape the public agenda, moderation of these spaces is both important and highly contested. Given this reality, we aim to explore which factors, alongside PME1 and PME3, impact support for social media content moderation. Because support for content moderation, as well as free speech concerns tethered to content moderation by platforms, are related to political ideology (Ballard, 2019), we include as variables political attitudes and support for free speech. Other domains of interest include general attitudes and behaviors toward platforms that may impact support for content moderation, such as frequency of social media use, pre-existing trust in platforms, ethical evaluations of platforms, and beliefs about whether platforms intentionally censor viewpoints. Anticipating that demographics will also factor into support for social media content moderation, we ask the following:

RQ1. What are the strongest predictors of support for social media content moderation?

Platforms and support for governmental regulation

In the United States, 46% of the internet population gets news from social media (Newman et al., 2019). According to the Pew Research Center (2019), 70% of Americans use social media, a drastic increase from 5% in 2005. The term *platform* describes a technology corporation offering services and communication infrastructure to users (Gillespie, 2010; Helmond, 2015). Platforms depend upon attention and engagement, both of which are crucial for advertising. When platforms create offerings that impact how the internet evolves, they become the infrastructure on which others rely. As such, platform users are inevitably "susceptible to processes of capitalisation and proprietary enclosure" (Mackenzie, 2019, p. 2003). As Helmond (2015) argues, the key determinants of platforms are programmability and structuring through application programming interfaces (APIs).

Scholars regularly contemplate the affordances that social platforms provide their users. In this vein, *imagined affordance* is a "term that helps scholars to reflect technological environments' material qualities that mediate affective experiences" (Nagy & Neff, 2015, p. 2). In this vein, we are interested in how users perceive the benign effects of social media platforms. While literature typically suggests stronger effects on others than on the self, we anticipate the opposite – a first-person perception – would be the case for beneficial effects of platforms. We argue this might be for three reasons: First, in line with the bias the Dunning-Kruger effect posits – namely that people overestimate their own cognitive skills (Dunning, 2011) – we suggest that one's individual relationship to technology and the direct benefits expected from technology might be easier to grasp and heuristically process than the role of technology and its beneficial effects on society writ large. Second, because technology conveys the myth of a steady increase in

convenience for users (Postman, 1993), we assume that presumed influence on self will be stronger – people perceiving social media platforms as having a bigger beneficial effect – than presumed influence on others. Third, we posit that the proposition of social desirability of media content leading to first-person effects (Golan & Day, 2008) extends beyond the context of content and is also applicable to platforms themselves. A Pew Research Center survey points into this direction: While 36% of Americans say that tech companies impact on society has been more bad than good, 63% think the opposite (Smith, 2018). And while 24% think the impact on the self has been more bad than good, 74% believe the opposite (Smith, 2018). Therefore, we posit the following hypothesis:

> **H2.** Participants will perceive stronger beneficial effects of social media platforms on themselves than on society.

Research underscores how social media platforms facilitate hierarchical relationships that shape much of the Internet as well as the rest of society (Fuchs, 2017). Case in point: In 2019 Facebook announced political actors would be allowed to advertise falsehoods and therefore held to different standards with regard to truth and speech norms than other users (Kreiss & McGregor, 2019). Asymmetries in how content is moderated as well as beliefs about social media platforms may not only exacerbate perceptions on how the self and others are affected, but also shape attitudes in favor of government regulation and/or content moderation.

Pew data from 2018 shows that 51% of Americans thought that "major technology companies (…) should be regulated more than they are now" (Smith, 2018, p. 7). We want to illuminate how different factors may impact regulation of social media platforms, including presumed influence on self (PME1) and on others (PME3), political attitudes, attitudes toward platforms, and demographic variables. To that end, we ask:

> **RQ2.** What are the strongest predictors of support for government regulation of platforms?

Methods

The data used in this study were collected through a cross-sectional U.S. national panel survey conducted by the Digital Media Research Program at the University of Texas at Austin. The survey was administered online using the survey software package Qualtrics. This study received Institutional Review Board approval on January 30, 2019, and the survey data were collected in March and April of the same year. Respondents were recruited by Dynata, an international survey company that provides access to panels that represent the adult online population of the United States.

Online panel data carry limitations, some of which we sought to overcome by implementing quotas based on gender, age, and race/ethnicity to match the distribution of these characteristics in the adult internet population in the United States as reported in December 2018 by the Pew Research Center. Previous research has validated this technique (e.g., Bode et al., 2014; Kim & Chen, 2015). The quota sampling process continued until subgroups (in our case, gender, age, race/ethnicity, as well as Hispanic yes/no) were fully populated and reached their respective

The Version of Record of this manuscript has been published and is available in
Information, Communication & Society (2021), https://doi.org/10.1080/1369118X.2021.1874040

quotas. In total, 1,465 people responded to the survey (after sorting out unfinished cases), and 443 further cases were screened out or excluded for failing quality checks implemented in the data collection process, such as speeding, straight-lining, or failed attention checks[i]. We implemented rigorous rules for screen-outs that are in line with the Pew Research Center's most recent recommendations for identifying bogus survey takers (Kennedy et al., 2020). The final respondents to the survey (N = 1,022) were slightly more female, less Hispanic, more educated and wealthier than the U.S. internet population (see Table 1).

Table 1
Demographic Profile of the U.S. Survey and U.S. Census

	Authors' Study, U.S. Survey, March 2019	Pew Research Center, U.S. Survey, December 2018
	(%)	(%)
Age:		
18-29	23.3	25.0
30-49	38.4	36.0
50-64	26.5	26.0
65+	11.8	12.0
Gender:		
Male	46.8	49.0
Female	53.2	51.0
Race/Ethnicity:		
White	75.0	73.0
Black	11.4	12.0
Other	13.6	14.0
Hispanic:		
Yes	5.8	15.0
Education:		
High school or less	24.3	34.0
Some college	35.1	34.0
College+	40.6	32.0
Household Income:		
Less than $30K	27.1	31.0
$30K-50K	17.7	18.0
$50K-75K	20.6	14.0
Greater than $75K	34.6	37.0

Variables of interest

Dependent variables related to regulation

Support for social media content moderation was operationalized by using and adapting measures from Hoffner et al. (1999), originally conceived for censorship of television violence, to social media platforms. Via Likert-type items on a scale from 1 (strongly disagree) to 10 (strongly agree), we asked participants to rate three statements: *"*I support social media platforms prohibiting the publishing of certain kinds of content,*"* "Platforms should have review systems for all social media content before it is allowed to be published," and "Platforms should have review systems for all social media content after it is published" (n = 1,021, M = 6.23, SD = 2.26). These were used to form an index (α = .823).

Support for government regulation was operationalized using a single-item Likert-type scale ranging from 1 (strongly disagree) to 10 (strongly agree), prompting respondents to evaluate the statement "Government should regulate social media platforms more than they are regulated now" (Smith, 2018), n = 1,022, *M* = 4.92, *SD* = 2.77.

Perceived effects on self and others

Perceived effects of social media content on oneself builds on measures by Chen and Ng (2017). This variable was operationalized using a Likert-type scale ranging from 1 (not at all) to 10 (very much), prompting respondents to evaluate three statements: "How much do you think you are influenced by content on social media platforms?," "How much do you think social media content leads you to be angry?," and "How much do you think social media content leads you to be upset?" These were averaged into an index, n = 1,017, *M* = 4.81, *SD* = 2.37, α = .855.

Perceived effects of social media content on others followed the same wording as the previous variable, but instead of 'you,' the three items asked how 'other adults' were impacted by social media content. This is in analogy to the dichotomy of 'you' vs. 'other people' used in TPE research (e.g., Hoffner et al., 1999). The resulting index had good reliability, n = 1,018, *M* = 7.14, *SD* = 1.86, *α* = .882.

Perceived impact of social media platforms on self. On a Likert-type scale from 1 (strongly disagree) to 10 (strongly agree), respondents were prompted to evaluate the statement "The impact that social media platforms and their products and services have had on me personally has been more good than bad" (Smith, 2018), n = 1,022, *M* = 5.23, *SD* = 2.452.

Perceived impact of social media platforms on society. On a Likert-type scale from 1 (strongly disagree) to 10 (strongly agree), respondents were prompted to evaluate the statement "The impact that social media platforms have had on society has been more good than bad" (Smith, 2018), n = 1,022, *M* = 5.04, *SD* = 2.420.

Political attitudes

Partisanship was operationalized through a Likert-type scale ranging from 0 (Strong Republican) over 5 (Independent) to 10 (Strong Democrat), after the prompt: "Generally speaking, do you usually think of yourself as a Republican, a Democrat, or an Independent?", n = 1,022, *M* = 6.34, *SD* = 3.04.

Opposition to censorship was operationalized using a subscale with good reliability (Alvarez & Kemmelmeier, 2018). We used a Likert-type scale ranging from 1 (strongly disagree) to 10 (strongly agree), for eight statements including "It is better to limit some violent or offensive speech than to allow all of it" (reverse-coded), "All points of view, no matter how offensive, should be allowed to be expressed in public (e.g., at rallies, public demonstrations, protests, etc.)," or "If it causes severe distress on others, public speech should be heavily restricted" (reverse-coded). Based on this, we formed an index, *n* = 1,021, *M* = 5.34, *SD* = 2.05, *α* = .888.

Support for free speech was operationalized based on censorship measures used by Rojas, Shah and Faber (1996). We used a Likert-type scale ranging from 1 (strongly disagree) to 10 (strongly agree) for three statements: "No matter how controversial an idea is, an individual should be able to express it publicly," "Everybody should have full liberty of promoting what they believe to be true," and "All individuals should have the right to openly express their ideas, no matter how prejudiced they might be," for which we recoded one item before forming an index with acceptable reliability, n = 1,022, M = 6.91, SD = 2.16, α = .860.

Platform attitudes

Perceived intentional censorship was operationalized using a Likert-type scale ranging from 1 (strongly disagree) to 10 (strongly agree), prompting respondents to evaluate the statements "Social media platforms intentionally censor viewpoints,' "The people working at social media platforms intentionally censor viewpoints," and "Algorithms at social media platforms intentionally censor viewpoints" (Smith, 2018). These were averaged into an index, n = 1,020, M = 6.10, SD = 2.20, α = .876.

Frequency of social media use was operationalized using a single-item Likert-type scale ranging from 1 (rarely/never) to 10 (often), which asked participants "How often do you use social media sites?" (Correa et al., 2010), n = 1,018, M = 6.80, SD = 3.26.

Trust in social media platforms was operationalized using a single-item Likert-type scale ranging from 1 (strongly disagree) to 10 (strongly agree), prompting respondents to evaluate the statement "I can trust social media platforms to do what is right" (Smith, 2018), n = 1,022, M = 4.09, SD = 2.48.

Perceived ethical behavior of platforms was operationalized using a single-item Likert-type scale ranging from 1 (strongly disagree) to 10 (strongly agree), prompting respondents to evaluate the statement "Social media platforms are ethical" (Smith, 2018), n = 1,022, M = 4.70, SD = 2.35.

Demographics

We sought to control for four relevant demographic variables: age (M = 44.47, SD = 16.20), gender, race, whether someone was Hispanic, and education (see Table 1 for demographics). All variables except age were dummy-coded for subsequent analyses.

Results

In line with research on the third-person effect, *H1* predicted that participants would perceive the effects of social media content on others to be stronger than on themselves. We ran a paired t-test and found a significant difference between effects on oneself (M = 4.81, SD = 2.37) and others (M = 7.13, SD = 1.86); t(1016)=-28.380, p =.000. We used Hierarchical Ordinary Least Squares (OLS) regression to answer *RQ1*, which asked what the strongest predictors of support for social media content review were. When all variables were entered in the full model, they accounted for 40.7% of the variance in the dependent variable, support for social media content review ($R^2_{Adjusted}$ = .407, F(15, 996) = 47.262, p < .001). Based on standardized beta coefficients (Table

2), we found that age (β = .098, $p < .001$), a higher level of education (β = .075, $p < .05$), opposition to censorship (β = -.553, $p < .001$) and the perceived effects of social media content on others (β = .161, $p < .001$) had significant effects on support for social media content review.[ii] In the tradition of the diamond model, we follow Schmierbach et al.'s (2011) recommendation of conducting a first regression with PME1 and PME3 as separate independent variables, and a second regression with the self and other as a summative term for second-person perceptions. While our results and the discussion section primarily focus on the first regression, we also report results from the diamond model ($R^2_{Adjusted}$ = .402, $F(14, 997) = 49.603$, $p < .001$) in Table 2.

Table 2
Support for social media content moderation hierarchical OLS regression

	Model 1: Perceived effects on self and others as predictors (Final model)		Model 2: Second-person perception as predictor (Final model)	
	b	*β*	*b*	*β*
Demographics				
Age	.014***	.098	.016***	.113
	(.004)		(.004)	
Gender (Male)	-.087	-.019	-.074	-.016
	(.117)		(.117)	
Race (White)	.257	.049	.285*	.055
	(.135)		(.136)	
Hispanic (Non-hispanic)	.282	.029	.303	.031
	(.241)		(.242)	
Education: Some college (high school or less)	.130	.027	.134	.028
	(.146)		(.146)	
Education: College plus (high school or less)	.348*	.075	.352*	.076
	(.144)		(.145)	
Adjusted R^2	4.2%		4.2%	
Political attitudes				
Partisanship	.029	.039	.028	.037
	(.020)		(.020)	
Opposition to censorship	-.612***	-.553	-.620***	-.560
	(.031)		(.031)	
Free speech support	.043	.041	.058*	.055
	(.028)		(.028)	
R^2 Change	33.8%		33.8%	
Platform attitudes and use				
Perceived intentional censorship	-.005	-.005	.002	.002
	(.028)		(.028)	
Frequency of social media use	-.001	-.002	-.003	-.004
	(.019)		(0.19)	
Trust in platforms	-.011	-.012	-.024	-.026
	(.033)		(.032)	
Perceived ethical behavior of platforms	.060	.062	.046	.047
	(.034)		(.034)	
R^2 Change	0.3%		0.3%	
Perceived effects on self/other/ self+other				
Perceived effects of social media content on self	.052	.055	-	-
	(.027)			

Perceived effects of social media content on others	.197*** (.034)	.161	-	-
Perceived effects of social media content on self+other	-	-	.112*** (.018)	.166
R^2 Change	2.7%		2.2%	
Total adjusted R^2	40.7%		40.2%	

Note. Standard errors are shown in parentheses, n = 1,012.
*p < .05. **p < .01. ***p < .001.

H2 predicted participants would perceive the beneficial effects of social media platforms to be stronger on themselves than on society. A paired t-test between effects on oneself (*M* = 5.23, *SD* = 2.45) and on society (*M* = 5.04, *SD* = 2.42) confirmed this hypothesis; t(2021)=-2.805, *p* = .005.

OLS regression helped us answer *RQ2*, which examined the predictors of support for government regulation of social media platforms. When all variables were entered into the full model, they accounted for 16.1% of the variance in the dependent variable, support for social media content review ($R^2_{Adjusted}$ = .161, $F(15, 1000) = 14.022$, $p < .001$). In analogy to the content-based regression, we also ran a regression for the summative term suggested by the diamond model ($R^2_{Adjusted}$ = .160, $F(14, 1001) = 14.829$, $p < .001$), reported in Table 3. The standardized beta coefficients (Table 3) display effect sizes in relation to each other. In the first regression with PME1 and PME3, we found that, among political attitudes, opposition to censorship (β = -.231, $p < .001$) had a significant negative effect on support for government regulation of platforms. Among platform attitudes, perceived intentional censorship (β = .251, $p < .001$), frequency of social media use (β = -.187, $p < .001$), and trust in platforms (β =.188, $p < .001$) had significant effects. Neither demographics nor PME1 or PME3 had significant effects.

Table 3
Support for platform regulation hierarchical OLS regression

	Model 1: Perceived effects on self and society as predictors (Final model)		Model 2: Second-person perception as predictor (Final model)	
	b	*β*	*b*	*β*
Demographics				
Age	.003 (.005)	.017	.003 (.005)	.015
Gender (Male)	-.193 (.169)	-.035	-.195 (.169)	-.035
Race (White)	.371 (.196)	.058	.361 (.196)	.057
Hispanic (Non-hispanic)	-.654 (.349)	-.055	-.642 (.350)	-.054
Education: Some college (high school or less)	-.291 (.211)	-.050	-.283 (.211)	-.049
Education: College plus (high school or less)	-.235 (.209)	-.042	-.230 (.209)	-.041
Adjusted R^2	1.1%		1.7%	
Political attitudes				
Partisanship	-.007	-.008	-.006	-.006

	(.028)		(.028)	
Opposition to censorship	-.311***	-.231	-.309***	-.229
	(.044)		(.044)	
Free speech support	.020	.016	.011	.008
	(.041)		(.040)	
R^2 Change		6.3%		6.3%
Platform attitudes and use				
Perceived intentional censorship	.317***	.251	.322***	.256
	(.039)		(.039)	
Frequency of social media use	-.159***	-.187	-.154***	-.181
	(.028)		(.028)	
Trust in platforms	.209***	.188	.202***	.182
	(.049)		(.049)	
Perceived ethical behavior of platforms	.005	.004	.005	.004
	(.051)		(.051)	
R^2 Change		9.2%		9.2%
Perceived effects on self/society/ self+society				
Perceived effects of platforms on self	.051	.045	-	-
	(.045)			
Perceived effects of platforms on society	-.071	-.062	-	-
	(.048)			
Perceived effects of platforms on self+society	-	-	-.008	-.012
			(.025)	
R^2 Change		0.2%		0.0%
Total adjusted R^2		16.1%		16.0%

Note. Standard errors are shown in parentheses, n = 1,016.
*p < .05. **p < .01. ***p < .001.

Discussion

The purpose of this study was to measure public attitudes regarding two important forms of social media regulation in society: content moderation *through* social media platforms and government regulation *of* social media platforms. We explored presumed negative effects of social media content on self and others and presumed beneficial effects of platforms on society and on the self, as well as how both perceptions relate to support for content moderation and platform regulation, respectively. We identified third-person perceptions regarding the presumed negative effects of social media content, and first-person perceptions regarding the presumed positive effects of platforms on society.

In our model, significant predictors of support for social media content moderation included age, education, opposition to censorship, and perceived effects of social media content on others (PME3). It stands to reason that opposition to censorship surfaces as a strong negative predictor of support for content moderation. Research has documented severe concerns about freedom of expression in light of content moderation (Ballard, 2019). When someone is fundamentally opposed to censorship, their opposition appears to extend into the realm of moderation. Our study aligns with what scholarship on PME1 and PME3 predicts (Chung & Moon, 2016). In the domain of social media content and its moderation, we find that PME3 significantly impacts support for moderation, while PME1 does not. When additionally consulting the diamond

method, we found a significant effect of the summative term of PME1 + PME3 on support for moderation.

Moving on to support for government regulation of platforms, we found a first-person perception: The perceived beneficial effects of social media platforms were significantly stronger on oneself than on society. Research suggests that first-person perceptions may occur in the context of desirable content (Golan & Day, 2008). Our research extends the applicability of a first-person perspective to communicative infrastructures – social media platforms. We can only speculate why that might be the case. The imagined affordances of platforms (Nagy & Neff, 2015) may be easier to conceptualize on a personal level that is rooted in one's own experience of platforms vis-à-vis the abstract notion of beneficial effects of platforms on society. While we did not measure underlying psychological mechanisms, the proposition that self-enhancement (Meirick, 2005) may be governing how the benefits of social media platforms were felt more strongly for the self than for society provides a compelling possible explanation.

The full regression model analyzing support for government regulation of platforms did not find significant effects of PME1 or PME3 on support for government regulation of platforms. It is possible that the imaginable larger benefits of platforms are difficult to conceptualize after all, so much so that an articulation of support for regulation becomes tricky. Via the diamond model, we also did not find significant effects of the summative term of PME1 and PME3 on support for regulation.

Like in the content moderation model, the main regression on platform regulation surfaced opposition to censorship as a strong negative predictor. However, perceived intentional censorship had a strong positive effect. This is noteworthy: When users think that platforms are intentionally censoring viewpoints, they are more supportive of government regulation. Shifts in public sentiment about how social media companies regulate content may be conducive to increased future support for government intervention in and regulation of technology companies.

Our results highlight the importance of attitudinal research about the role of social media platforms in society, particularly about what users think are possible harms and benefits of social media content and platforms. Platform operators and lawmakers must take seriously the constituents of platforms when setting up regimes of governance so that proportionality of measures is maintained. Studies such as ours help gauge public opinion on how to understand the powers and impact of social media platforms – knowledge which, as we argue, is critical in shaping and defining policy.

This study expands research on the third-person effect by confirming the predictive power of PME3 for the behavioral variable of support for content moderation, but not for support for platform regulation. This is mirrored by results from the diamond model. We identify third-person perceptions with regard to the impact of social media content, and first-person perceptions with regard to the impact of platforms writ large, thus extending the applicability of research on first-person perceptions from content toward infrastructures such as social media platforms.

Limitations

This study is not without limitations. The survey design prohibited us from exploring further the motivations for specific perceptions. Furthermore, as a cross-sectional study, we cannot make conclusive statements about causal relations. Measuring perceptions toward regulation of and by platforms was conducted with items about the negative impact of content on 'you' vs. 'other people,' and about the positive impact of platforms on 'me personally' vs. 'society.' While differences in measurement are tied to how the survey study was conducted, the lack of parallel measurements posits a limitation to our study. Within the domain of TPP and FPP, measures typically distinguish between other individuals and the self, rather than between other individuals and society. This conceptual nuance was not maintained in our measures. We acknowledge this as a limitation to our study and invite future researchers to test these relationships with updated sets of items. We included the summative term of effects on others and self in the spirit of Schmierbach et al. (2011) for future research with a particular emphasis on second-person perceptions (Neuwirth & Frederick, 2002) to engage with. Exemplary research on the diamond method includes Sun, Shen and Pan's (2008) work which can be helpful in mapping paths forward for more research in this area.

In lieu of more sophisticated models (e.g., on interrelationships between frequency of social media use, exposure to negative content, and first-person perceptions), we opted for parsimonious models aimed at our primary research questions and hypotheses. While the use of single-item measures is not preferred from a psychometric standpoint, regardless of their face validity, in this study it allowed for the inclusion of a variety of measures in a large-scale survey (Scherr, 2018). We suggest future research to test and develop multiple-item scales alongside our single-item measures (Fuchs and Diamantopoulos, 2009).

References

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211

Allen, M., & VandeHei, J. (2019, September 1). Trump allies plot new war on social media. *Axios*. Retrieved from https://www.axios.com/trump-2020-campaign-social-media-bias-41bbed1e-0bd3-445d-bf6f-195b5c3e65a6.html

Alvarez, M. J., & Kemmelmeier, M. (2018). Free speech as a cultural value in the United States. *Journal of Social and Political Psychology*, *5*(2), 707–735. https://doi.org/10.5964/jspp.v5i2.590

Baek, Y. M., Kang, H., & Kim, S. (2019). Fake news should be regulated because it influences both "others" and "me": How and why the influence of presumed influence model should be extended. *Mass Communication and Society*, *22*(3), 301–323. https://doi.org/10.1080/15205436.2018.1562076

Ballard, J. (2019, April 29). Most conservatives believe removing content and comments on social media is suppressing free speech. *YouGov*. Retrieved from https://today.yougov.com/topics/technology/articles-reports/2019/04/29/content-moderation-social-media-free-speech-poll

Betts, L. R., Metwally, S. H., & Gardner, S. E. (2019). We are safe but you are not: Exploring comparative optimism and cyber bullying. *Journal of Technology in Behavioral Science*,

4(3), 227–233. https://doi.org/10.1007/s41347-018-0070-6

Bode, L., Vraga, E. K., Borah, P., & Shah, D. V. (2014). A new space for political behavior: Political social networking and its democratic consequences. *Journal of Computer-Mediated Communication*, *19*(3), 414–429. https://doi.org/10.1111/jcc4.12048

Chen, G. M. (2017). *Online incivility and public debate: Nasty talk*. Palgrave Macmillan.

Chen, A. (2017, January 28). The human toll of protecting the internet from the worst of humanity. Retrieved from http://www.newyorker.com/tech/elements/the-human-toll-of-protecting-the-internet-from-the-worst-of-humanity

Chen, G. M., & Ng, Y. M. M. (2016). Third-person perception of online comments: Civil ones persuade you more than me. *Computers in Human Behavior*, *55*, 736–742. https://doi.org/10.1016/j.chb.2015.10.014

Chen, G. M., & Ng, Y. M. M. (2017). Nasty online comments anger you more than me, but nice ones make me as happy as you. *Computers in Human Behavior*, *71*, 181–188. https://doi.org/10.1016/j.chb.2017.02.010

Cheng, Y., & Chen, Z. F. (2020). The influence of presumed fake news influence: Examining public support for corporate corrective response, media literacy interventions, and governmental regulation. *Mass Communication and Society*, *23*(5), 705–729. https://doi.org/10.1080/15205436.2020.1750656

Chia, S. C., Li, H., Detenber, B., & Lee, W. (2006). Mining the internet plateau: An exploration of the adoption intention of non-users in Singapore. *New Media and Society*, *8*(4), 589–609. https://doi.org/10.1177/1461444806065656

Chung, S., & Moon, S.-I. (2016). Is the Third-Person Effect real? A critical examination of rationales, testing methods, and previous findings on the Third-Person Effect on censorship attitudes. *Human Communication Research*, *42*(2), 312–337. https://doi.org/10.1111/hcre.12078

Correa, T., Hinsley, A., & de Zúñiga, H. G. (2010). Who interacts on the Web?: The intersection of users' personality and social media use. *Computers in Human Behavior*, *26*(2), 247–253. https://doi.org/10.1016/j.chb.2009.09.003

Davison, W. P. (1983). The third-person effect in communication. *Public Opinion Quarterly*, *47*(1), 1–15. https://doi.org/10.1086/268763

Dunning, D. (2011). The Dunning-Kruger effect: On being ignorant of one's own ignorance. In J. M. Olson & M. P. Zanna (Eds.), *Advances in experimental social psychology* (Vol. 44, pp. 247-296). Academic Press. https://doi.org/10.1016/B978-0-12-385522-0.00005-6

Fuchs, C. (2017). *Social media: A critical introduction*. Sage.

Fuchs, C., & Diamantopoulos, A. (2009). Using single-item measures for construct measurement in management research: Conceptual issues and application guidelines. *Die Betriebswirtschaft*, *69*(2), 195.

Gervais, B. T. (2015). Incivility online: Affective and behavioral reactions to uncivil political posts in a web-based experiment. *Journal of Information Technology and Politics*, *12*(2), 167–185. https://doi.org/10.1080/19331681.2014.997416

Gillespie, T. (2010). The politics of "platforms." *New Media and Society*, *12*(3), 347–364. https://doi.org/10.1177/1461444809342738

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Golan, G. J., & Day, A. G. (2008). The first-person effect and its behavioral consequences: A new trend in the twenty-five year history of third-person effect research. *Mass*

Communication and Society, *11*(4), 539–556. https://doi.org/10.1080/15205430802368621

Gorwa, R. (2019). What is platform governance? *Information, Communication and Society, 22*(6), 854–871. https://doi.org/10.1080/1369118X.2019.1573914

Grimmelmann, J. (2015). The virtues of moderation. *The Yale Journal of Law & Technology*, *17*(42), 42–109.

Gunther, A. (1991). What we think others think: Cause and consequence in the third-person effect. *Communication Research*, *18*(3), 355–372. https://doi.org/10.1177/009365091018003004

Gunther, A. C., & Mundy, P. (1993). Biased optimism and the third-person effect. *Journalism Quarterly, 70*(1), 58-67. https://doi.org/10.1177/107769909307000107

Gunther, A. C., & Storey, J. D. (2003). The influence of presumed influence. *Journal of Communication*, *53*(2), 199-215. https://doi.org/10.1111/j.1460-2466.2003.tb02586.x

Helmond, A. (2015). The platformization of the web: Making web data platform ready. *Social Media & Society*, 1(2), 1-11. https://doi.org/10.1177/2056305115603080

Hoffner, C., Buchanana, M., Anderson, J. D., Hubbs, L. A., Kamigaki, S. K., Kowalczyk, L., … Silberg, K. J. (1999). Support for censorship of television violence: The role of the third-person effect and news exposure. *Communication Research*, *26*(6), 726–742. https://doi.org/10.1177/009365099026006004

Jack, C. (2019). Wicked content. *Communication, Culture and Critique, 12*(4), 435–454. https://doi.org/10.1093/ccc/tcz043

Jang, S. M., & Kim, J. K. (2018). Third person effects of fake news: Fake news regulation and media literacy interventions. *Computers in Human Behavior*, *80*, 295–302. https://doi.org/10.1016/j.chb.2017.11.034

Kennedy, C., Hatley, N., Lau, A., Mercer, A., Keeter, S., Ferno, J., & Asare-Marfo, D. (2020). Assessing the risks to online polls from bogus respondents. *Pew Research Center.* Retrieved from https://www.pewresearch.org/methods/2020/02/18/assessing-the-risks-to-online-polls-from-bogus-respondents/

Kim, Y., & Chen, H. T. (2015). Discussion network heterogeneity matters: Examining a moderated mediation model of social media use and civic engagement. *International Journal of Communication*, *9*(1), 2344–2365. https://ijoc.org/index.php/ijoc/article/view/3254

Klonick, K. (2017). The new governors: The people, rules, and processes governing online speech. *Harvard Law Review*, *131*, 1598–1670.

Kreiss, D., & McGregor, S. C. (2019). The "Arbiters of What Our Voters See": Facebook and Google's Struggle with Policy, Process, and Enforcement around Political Advertising. *Political Communication*, *36*(4), 499-522. https://doi.org/10.1080/10584609.2019.1619639

Lee, B., & Tamborini, R. (2005). Third-person effect and internet pornography: The influence of collectivism and internet self-efficacy. *Journal of Communication*, *55*(2), 292–310. https://doi.org/10.1111/j.1460-2466.2005.tb02673.x

Lee-Won, R. J., White, T. N., Song, H., Lee, J. Y., & Smith, M. R. (2019). Source magnification of cyberhate: affective and cognitive effects of multiple-source hate messages on target group members. *Media Psychology*, 1-22. https://doi.org/10.1080/15213269.2019.1612760

Lo, V.-H., & Wei, R. (2002). Third-person effect, gender, and pornography on the internet.

Journal of Broadcasting and Electronic Media, *46*(1), 13–33. https://doi.org/10.1207/s15506878jobem4601_2

Mackenzie, A. (2019). From API to AI: platforms and their opacities. *Information, Communication and Society*, *22*(13), 1989–2006. https://doi.org/10.1080/1369118X.2018.1476569

Matias, J. N. (2019). The civic labor of volunteer moderators online. *Social Media and Society*, *5*(2), 1–12. https://doi.org/10.1177/2056305119836778

Medeiros, B. (2017). Platform (non-)intervention and the "marketplace" paradigm for speech regulation. *Social Media + Society*, *3*(1), 1-10. https://doi.org/10.1177/2056305117691997

Meirick, P. C. (2005). Self-enhancement motivation as a third variable in the relationship between First- and Third-Person Effects. *International Journal of Public Opinion Research*, *17*(4), 473–483. https://doi.org/10.1093/ijpor/edh077

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media and Society*, *20*(11), 4366–4383. https://doi.org/10.1177/1461444818773059

Naab, T. K., Kalch, A., & Meitz, T. G. K. (2018). Flagging uncivil user comments: Effects of intervention information, type of victim, and response comments on bystander behavior. *New Media and Society*, *20*(2), 777–795. https://doi.org/10.1177/1461444816670923

Nagy, P., & Neff, G. (2015). Imagined affordance: Reconstructing a keyword for communication theory. *Social Media and Society*, *1*(2), 1–9. https://doi.org/10.1177/2056305115603385

Nan, X. (2007). Social distance, framing, and judgment: A construal level perspective. *Human Communication Research*, *33*(4), 489-514. https://doi.org/10.1111/j.1468-2958.2007.00309.x

Neuwirth, K., & Frederick, E. (2002). Extending the framework of Third-, First- and Second-Person Effects. *Mass Communication and Society*, *5*(2), 207–228. https://doi.org/10.1207/S15327825MCS0502

Newman, N., Fletcher, R., Kalogeropoulos, A., & Nielsen, R. K. (2019). *Reuters Institute Digital News Report 2019*. *Reuters Institute for the Study of Journalism, University of Oxford.* Retrieved from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2019-06/DNR_2019_FINAL_0.pdf

Paradise, A., & Sullivan, M. (2012). (In)visible threats? The third-person effect in perceptions of the influence of Facebook. *Cyberpsychology, Behavior, and Social Networking*, *15*(1), 55–60. https://doi.org/10.1089/cyber.2011.0054

Pew Research Center. (2019, June 12). Social media fact sheet. Retrieved from https://www.pewresearch.org/internet/fact-sheet/social-media/

Pohjonen, M. (2019). A comparative approach to social media extreme speech: Online hate speech as media commentary. *International Journal of Communication*, *13*, 3088-3103. https://ijoc.org/index.php/ijoc/article/view/9110

Postman, N. (1993). *Technopoly: The surrender of culture to technology.* Vintage Books

Riedl, M. J., Masullo, G. M., & Whipple, K. N. (2020). The downsides of digital labor: Exploring the toll incivility takes on online comment moderators. *Computers in Human Behavior, 107.* https://doi.org/10.1016/j.chb.2020.106262

Roberts, S. T. (2019). *Behind the screen: Content moderation in the shadows of social media*. Yale University Press.

Rojas, H., Shah, D. V, & Faber, R. J. (1996). For the good of others: Censorship and the third-

person effect. *International Journal of Public Opinion Research*, *8*(2), 163–186. https://doi.org/10.1093/ijpor/8.2.163

Salwen, M. B., & Dupagne, M. (1999). The third-person effect: Perceptions of the media's influence and immoral consequences. *Communication Research*, *26*(5), 523–549. https://doi.org/10.1177/009365099026005001

Scherr, S. (2018). Traditional media use and depression in the general population: evidence for a non-linear relationship. *Current Psychology*. https://doi.org/10.1007/s12144-018-0020-7

Schmierbach, M., Boyle, M. P., Xu, Q., & McLeod, D. M. (2011). Exploring Third-Person differences between gamers and nongamers. *Journal of Communication*, *61*(2), 307–327. https://doi.org/10.1111/j.1460-2466.2011.01541.x

Smith, A. (2018, June 28). Public attitudes toward technology companies. *Pew Research Center*. Retrieved from https://www.pewinternet.org/2018/06/28/public-attitudes-toward-technology-companies/

Sun, Y., Shen, L., & Pan, Z. (2008). On the behavioral component of the Third-Person Effect. *Communication Research*, *35*(2), 257–278. https://doi.org/10.1177/0093650207313167

Tsay-Vogel, M. (2016). Me versus them: Third-person effects among Facebook users. *New Media and Society*, *18*(9), 1956–1972. https://doi.org/10.1177/1461444815573476

Woolley, S. C., & Howard, P. N. (Eds.). (2018). *Computational propaganda: political parties, politicians, and political manipulation on social media*. Oxford University Press.

Endnotes

[i] A relatively high rate of screen-outs can be explained by both our conservative quality criteria to protect the integrity of the collected dataset, as well as the length of the survey. Beyond what is reported, the survey also included dimensions on fake news, representation, (social) media and journalism. We carefully designed and pretested the questionnaire to avoid potential response bias. Few to no topically similar questions were asked prior to the ones used in this study, with the intent to avoid question order effects.

[ii] While we limited our analytical inquiry to PME3, PME1, and the effect of their summative term as suggested by the diamond method, we also calculated TPPs and entered them into the regression model. Neither in the content moderation nor the platform regulation model did TPP have a significant effect on the behavioral variables of support for regulation.