

2023

Proteus meets Eris? Understanding the influence of pseudonymous self-representation on instant messenger discussions

<https://hdl.handle.net/2144/45445>

Downloaded from DSpace Repository, DSpace Institution's institutional repository

BOSTON UNIVERSITY
COLLEGE OF COMMUNICATION

Dissertation

**PROTEUS MEETS ERIS?
UNDERSTANDING THE INFLUENCE OF PSEUDONYMOUS SELF-
REPRESENTATION ON INSTANT MESSENGER DISCUSSIONS**

by

ERIN ELIZABETH WERTZ

B.A., Rollins College, 2012
M.A., Boston University, 2016

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2023

Approved by

First Reader

James Cummings, Ph.D.
Assistant Professor of Emerging Media Studies

Second Reader

Lei Guo, Ph.D.
Associate Professor of Emerging Media Studies

Third Reader

Chris Wells, Ph.D.
Associate Professor of Emerging Media Studies

Fourth Reader

Virginia Sapiro, Ph.D.
Dean *Emerita* of the College and Graduate School of Arts and Sciences
Professor *Emerita* of Political Science

DEDICATION

To Lian: you aren't around to read this, but I wouldn't be who I am without you.

ACKNOWLEDGMENTS

First and foremost, I must acknowledge the tremendous effort and input my advisor Jim Cummings has invested in this project. Without him, this would be a far worse work, or perhaps nothing at all. The rest of the committee: professors Lei Guo, Chris Wells and Virginia Sapiro have also been wonderful mentors, and I greatly appreciate their endless patience with all the delays and difficulties this project has encountered.

I would also like to thank my parents for their love, support and encouragement throughout my PhD journey, as well as my friends Sabrina, NemoMarx, Squishy, Fish, K, gargulec and Chehrazad, who have invested a degree of faith and confidence in me that I find completely unearned.

PROTEUS MEETS ERIS?
UNDERSTANDING THE INFLUENCE OF PSEUDONYMOUS SELF-
REPRESENTATION ON INSTANT MESSENGER DISCUSSIONS

ERIN ELIZABETH WERTZ

Boston University College of Communication, 2023

Major Professor: James Cummings, Ph.D., Assistant Professor of Emerging Media
Studies

ABSTRACT

This dissertation proposes a novel definition of anonymity, drawing on past definitions as well as psychological theory, to propose that pseudonymous identities can have a complex and nuanced influence in emphasizing certain personality traits when used in online discussion. This dissertation connects this definition to the Proteus Effect — the observation that individuals adopt behavior stereotypical of the avatars they use in virtual worlds (Yee & Bailenson, 2007) — to test how the presence and character of avatars in an online instant messenger influences aggression during political discussions. A 2x2 factorial experiment is used to evaluate participant aggression following small group deliberations between groups of participants assigned aggressive and unaggressive usernames and avatars, as well as accounts displaying an avatar and username vs only a username. A follow-up online experiment is used to show that similar effects of identity on behavior can be achieved simply by assigning participants to participate in similar tasks as moderators or as themselves.

TABLE OF CONTENTS

DEDICATION	iv
ACKNOWLEDGMENTS	v
ABSTRACT.....	vi
TABLE OF CONTENTS.....	vii
LIST OF TABLES	x
LIST OF FIGURES	xi
LIST OF ABBREVIATIONS.....	xii
Introduction.....	1
Chapter 1: Anonymity in Online Political Talk.....	9
The Problems with “Trade-off” Anonymity	10
Past Attempts to Conceptually Define Anonymity	14
Anonymity in Theories of Deindividuation and Disinhibition	20
Anonymity as Social Identity in The Social Identity Model of Deindividuation Effects	20
Antecedents of Online Disinhibition	21
Anonymity as Identity Performance	25
Redefining Perceived Computer-Mediated Pseudonymous Anonymity	28
Perceived Anonymity	28
Self -Anonymity	29
Other-Anonymity.....	30
Summary	32
Chapter 2: The Proteus Effect.....	34
Causes of the Proteus Effect	35
The Proteus Effect and SIDE.....	37
The Proteus Effect and Online Disinhibition	40
Modality as a Moderating Factor in the Proteus Effect	42
The Complicating Role of Other-Anonymity	45

Chapter 3: Methodology	47
Participants	48
Materials	50
Avatar and usernames designs	50
Sampling and Pretesting of Stimuli.	51
Procedure	54
Measures	56
Self-Report Measures.	56
Behavioral Measures	60
Data Analysis Plan	63
Handling of Multiple Samples	64
Chapter 4: Results	66
Behavioral Outcomes	66
Endorsement of Aggressive Norms	71
Group Cohesion (c)	73
Alternative Model Specifications	75
Behavioral Outcomes	76
Endorsement of Aggressive Norms	80
Group Cohesion	82
Chapter 5: Discussion	85
Interpretation of Results	86
Implications	87
The Proteus Effect Occurs Through Avatar Perceptions	88
The Proteus Effect is not Dependent on a Particular Medium	89
Technology Moderates the Proteus Effect	91
Toxic Avatars Can Lead to Toxic Disinhibition	92
Limitations and Future Research	93
Chapter 6: Qualitative Assessment of Participant Discussions	97
Methodology	100

Results	103
Law and Order	103
Community Good	105
Education/Rehabilitation	107
Victims versus Oppressors	108
Free Speech.....	109
Discussion	110
Results and Implications.....	111
Limitations and Next Steps.....	113
Chapter 7: A Follow-Up Study	115
Methods.....	117
Participants	117
Materials	119
Results	120
Discussion	126
Implications	127
Limitations and Future Directions	128
Chapter 8: Conclusions	130
Anonymity Can Consolidate Existing Theory	133
Anonymity Need Not Be a Trade-Off.....	135
Anonymity is Many and Varied	137
Appendix A: Discussion Group Script	139
Appendix B: Study 1 Questionnaire	142
Appendix C: Alternate Multilevel Model Specifications	146
Bibliography	152
Curriculum Vitae	161

LIST OF TABLES

Table 1: Multilevel Models for Agreement	68
Table 2: Multilevel Models for Disagreement.....	69
Table 3: Multilevel Models for Elaboration	70
Table 4: Multilevel Models for Endorsement of Aggressive Norms.....	72
Table 5: Multilevel Models for Group Cohesion.....	74
Table 6: Regression Results for Agreement	77
Table 7: Regression Results for Disagreement.....	78
Table 8: Regression Results for Elaboration	79
Table 9: Regression Results for Endorsement of Aggressive Norms.....	81
Table 10: Regression Results for Group Cohesion.....	83
Table 11: Alternate Multilevel Models for Agreement	147
Table 12: Alternate Multilevel Models for Disagreement.....	148
Table 13: Alternate Multilevel Models for Elaboration	149
Table 14: Alternate Multilevel Models for Endorsement of Aggressive Norms.....	150
Table 15: Alternate Multilevel Models for Group Cohesion.....	151

LIST OF FIGURES

Figure 1: Aggressive (left) and unaggressive (right) avatars and usernames	54
Figure 2: Rankings for "No Punishment" Option	122
Figure 3: Rankings for "Apology" Option.....	123
Figure 4: Rankings for "Warning" Option.....	123
Figure 6: Rankings for "Permanent Ban" Option	124
Figure 5: Rankings for "One Week Ban" Option	124

LIST OF ABBREVIATIONS

AIC	Akaike Information Criterion
CMC	Computer-Mediated Communication
H#	Hypothesis # (e.g., 2, 1a)
ICC	Intraclass Correlation Coefficient
IRL	In Real Life
SIDE	Social Identity (Model) of Deindividuation Effects

Introduction

From early utopian dreams to 2010s fears of QAnon, fake news, and disinformation, much attention has been devoted to the notion of political talk online and its potential to influence broader democratic trends. Amidst credible accusations that Facebook's inability or unwillingness to moderate propaganda has facilitated a genocide in Myanmar (Mozer, 2018), that Reddit's free speech policies supported the notorious antifeminist harassment campaigns of #Gamergate, and that Twitter enabled the spread of covid misinformation (Rosenberg et al. 2020); there is much to be said that is negative about social media and democracy. At the same time, social media also been suggested to have facilitated the peaceful overthrow of dictators during the Arab Spring, (Howard et al., 2011), provided a critical outlet for LGBT+ and other marginalized voices to speak out (Haimson et al., 2021) and created a means for women to spread messages against sexual abuse in the #metoo campaign (Manikonda et al., 2018). Of course, all of these have had larger and more intricate causes than the media environment in which they happened. Nonetheless, the record on social media and civic society is somewhat mixed.

While the internet has clearly not met the lofty standards dreamt of by utopian deliberative theorists (e.g., Rheingold, 1993), there is much still to be learned about online civic participation as it actually happens online beyond utopian dreams. Wright (2012) argues that those studying the impact of the internet on democracy should abandon broader questions about how the internet will save and/or destroy and/or fail to change the world. Instead, Wright (2012) calls for more emphasis to be given to understanding both everyday political talk, and experimental research that attempts to

understand what factors can influence discussion. Everyday political talk is talk that is not necessarily purposive or deliberative, but which broaches on ideological political topics incidentally as individuals discuss reality TV or personal talk or provide advice on smaller scale problems within their communities (Connover & Searing, 2005).

Online there is a tremendous volume of this political talk, much of which contains sophisticated and cross-cutting dialogue (Wright et al., 2016). In discussions surrounding this genre of online talk that takes place in hobbyist forums and subreddits and Discord servers, few factors have attracted quite as much attention as anonymity. On the internet, where “no one knows you’re a dog,” (Steiner, 1993), the ability to adopt alternate identities or avoid disclosing any identity at all has been used by both utopians and cynics to make profound claims about the power of the internet to save and/or ruin political discourse (e.g., Kabay, 1993). However, anonymity is often poorly conceived both in public and academic debates (Moore et al., 2020) and, in the context of political discourse, much remains to be learned about how different forms of anonymity influence debate to different ends.

This dissertation attempts to take up Wright’s (2012) call to adopt a more experimental approach to the study of, particularly, everyday political talk online and examine the complex and divergent roles anonymity can play in influencing this talk. Anonymity, it is argued, has been as misunderstood as the internet and democracy itself, left undertheorized even as it has been credited and blamed with a staggering variety of consequences (Moore, 2020). To step past these sweeping claims and build a picture of multiple distinct ways of being anonymous, each with their own effects, this dissertation

first examines how empirical work has operationalized anonymity, as well as past attempts to define and theorize its effects. In doing so, anonymity is defined here not only to the act of hiding identity information, but also to the presentation of specific identity cues. This conceptualization of anonymity is, in turn, then connected to the media effects theory of the Proteus Effect, the finding that individuals adopt traits from the avatars they embody in virtual worlds (Yee & Bailenson, 2007). An experiment analyzing the divergent effects different pseudonymous identities can cause in influencing the tone of political talk online, as well as two follow-up projects on this same theme, are described. The following paragraphs present a more in-depth summary of each chapter of this dissertation:

Chapter 1: Anonymity in Online Political Talk first analyzes the idea that anonymity can be understood as a “trade-off” where allowing its benefits means accepting its drawbacks, first looking at the research that underpins this widespread view, and showing that it often fails to adequately theorize the concept of anonymity. To address this, this chapter also reviews prior conceptual efforts to define anonymity as well as theories of disinhibition that have been used to justify the argument that anonymity can change behavior. Drawing on the work of Postmes, Spears and Lea (1998) as well as Suler (2004), it is argued that these theories conceptualize anonymity not purely through hiding information, but also through making visible other information, a notion that has been largely ignored in prior definitions of anonymity. One exception to this is Asenbaum’s (2018) work which describes anonymity as a performance that creates distinct social identities. A novel definition of anonymity is proposed attempting to

integrate all of the above into a useful framework for empirical research for the study of perceived online anonymity: *Perceived online anonymity is the totality of identity information individuals perceive to be presented and not presented in computer-mediated contexts, including elements of both perceived self-anonymity — the range and intensity of identity cues an individual presents in online contexts — and perceived other-anonymity — perceptions of others’ social presence.*

Chapter 2: The Proteus Effect and Perceived Online Anonymity connects this definition of anonymity to the study of the Proteus Effect. The history of the Proteus Effect is reviewed, as well as a number of causal mechanisms thought to cause this effect. This chapter argues that the Proteus Effect, often relegated to specific three-dimensional virtual environments, fits within the prior redefinition of anonymity and should not be a priori constrained to virtual worlds specifically; rather, it can be taken to apply to a wide variety of media much less rich than virtual worlds. By juxtaposing the Proteus Effect and this definition of anonymity, it is hypothesized that individuals engaging in pseudonymous online political talk on instant messenger programs (e.g., WeChat, Discord, IRC) will likewise adopt avatar traits which may induce specific behavioral outcomes, namely increased aggression if those avatars are aggressive. In doing this, this study connects Chapter 1’s definition of anonymity to specific empirical tests which can validate its rebuttal of a “trade-off” conception of anonymity.

Chapter 3: Methods, Chapter 4: Results and Chapter 5: Discussion describe an online experiment using synchronous pseudonymous instant messenger chats with experimentally varied avatars and usernames to test the propositions generated in Chapter

2. Specifically, individuals were asked to envision themselves as moderators of an online community on Discord and to evaluate five options for punishing mild misbehavior within online communities, implicitly asking them to evaluate and prioritize ideals of free speech, community safety, and the value of tolerance in online spaces. Participants were assigned either matching aggressive (e.g., Dark Lurker) or unaggressive (e.g., Starboy) usernames. These usernames were either presented with both a username and matching avatar image displayed next to the messages they sent, or a username with no accompanying image. While, contrary to hypotheses, the direct experimental manipulations of account aggressiveness showed no effect on aggression, it was observed that subjective perceptions of account aggressiveness indeed led to increased endorsement of aggressive norms following the discussion and lower reported group cohesion of the discussion. These effects were stronger in the condition with avatars in addition to usernames than in the condition with usernames alone.

Chapter 6: Qualitative Assessment of Participant Discussions offers a more thorough qualitative examination of participant arguments in the study outlined in Chapters 3–5. Specifically, it follows an observation that I made while conducting the prior study that, while participants were asked to imagine themselves as moderators merely to get them to discuss their own perspectives on the issue of punishment in online communities, some participants seemed to internalize “moderator” as a salient identity label, in ways that seem to have guided their behavior and the arguments they made, as well as leading them to choose punishments which fulfilled the assumed responsibility of a moderator to act. This, it is argued, could have constituted a second, unanticipated, type

of Proteus Effect, whereby participants adopted not only the aggressive or unaggressive avatar traits of their assigned accounts but also the social identity of moderators. All instances where participants offered justification for their arguments were reexamined to identify instances of this behavior. As anticipated, the two most common types of justifications: Law and Order and Community Good, tended to reflect either explicit adoption of a moderator identity or implicit argument frames that grounded themselves in the identity of a moderator. This suggests that this second Proteus Effect did take place. However, without a stronger empirical grounding, it was impossible to verify the prevalence of this phenomenon or whether it was caused merely by assigning participants to the role of moderators, having them engage in extended discussions with others as moderators; or simply reflects how individuals would naturally approach the problem of moderation regardless of role assignment or discussion. To do this, a short follow-up study was designed.

Chapter 7: A Follow-Up Study describes the methods and results of this study. Participants were asked to complete the same task used in the previous discussion study by themselves, either while imagining themselves as a moderator or without any clear identity prime. These results were also compared to the rankings of punishment options participants in the prior study gave following group discussions. It was anticipated that participants asked to rank punishments as moderators would rank the punishments somewhat differently than those not asked to think of themselves as moderators, reflecting what the qualitative analysis observed as a potential Proteus Effect, with participants driven to answer the question as moderators instead of as they might

normally. Additionally, as the previous study's discussion task was hypothesized to reinforce the assigned identity of moderators, it was anticipated that this discussion would produce similar but stronger effects than those observed between the two groups merely asked to rank the punishments without discussion. No direct evidence of difference was found between the group assigned no identity and the group told to act as moderators. However, the group asked to discuss as moderators and then rank punishments (study 1 participants) did differ from both of the follow-up study groups, with chosen punishments being much more distinct from the group not assigned identities than with the group asked to rank punishments as moderators. These results are interpreted to provide support for a second Proteus Effect, drawing not on avatars, but merely on the assignment of a salient identity to participants. This, in turn, is argued to support interpreting the Proteus Effect as a specific case of the general notion of self-anonymity as discussed in Chapter 1.

Finally, **Chapter 8: Conclusions** summarizes the preceding sections and attempts to widen the conversation back to the larger topic of everyday online political talk. The entirety of the project is summarized, then implications of findings are discussed. Anonymity can serve to consolidate and contrast several theories of online identity that have framed discussion of anonymity as intrinsically disinhibiting. Anonymities should be considered as a varied and diverse set of phenomena, rather than as a necessary trade-off where all benefits and drawbacks must co-occur. While trying to engineer for specific anonymities is a somewhat risky proposition, the range of possible anonymities is nonetheless something that platform owners should understand and consider when

designing policy for political talk. This work helps to advance Wright's (2012) call to move past sweeping all or nothing perspectives when it comes to the internet and democracy.

Chapter 1: Anonymity in Online Political Talk

Anonymity in the study of online political talk, imagined as the hiding of personal information, particularly one's name from those with which one talks (Anonymous, 1998), is often presented as offering an inevitable mix of benefits and drawbacks (e.g., Strandberg & Grönlund, 2018; Friess & Eilders, 2015; Williams, 2005). Among other benefits, proponents have argued that anonymity reduces the deleterious influence of offline social status and identity (Chester & Gwynne, 1998), frees people from fear of reprisal for expressing opinions critical of repressive institutions (Jardine, 2015), lowers thresholds for participation in discussion (Cho & Acquisti, 2013) and makes it easier to adopt intimacy and hold frank conversation (Bernstein et al., 2011). By contrast, anonymity has also been blamed for antisocial behavior, incivility, harassment and the widespread promotion of racism (e.g., Rainie et al., 2017). Following the US 2016 presidential election, concerns about online disinformation and fake news, as well as the spread of misinformation have also been blamed in part on the widespread existence of anonymity online (Tucker et al., 2017). In such work, anonymity is generally understood as presenting a trade-off where accepting the benefits entails dealing with the drawbacks.

This perspective draws support from the online disinhibition effect (Suler, 2004) which argues that online communication provides a number of affordances which can trigger disinhibition. This perspective, necessarily, frames the issue of anonymity as a policy choice of whether to value the benefits or the drawbacks of anonymity. However, researchers have observed that anonymity itself is often undertheorized or examined in limiting and fragmenting terms (e.g., Moore, 2018; Scott & Raines, 2020; Clark-Gordon

et al., 2019) and that this perspective may be treating the underlying concept of anonymity too simply to adequately theorize its effects.

This chapter will first look as to how the above research has constructed (or failed to construct) anonymity in practice, before evaluating several more compelling theoretical accounts of anonymity that have rarely been operationalized adequately in the empirical literature. By pairing these accounts with the Social Identity Model of Deindividuation effects (SIDE Model; Postmes et al. 1997) and Suler (2004)'s online disinhibition effect, as well as recent advances in queer and feminist theory, this section will argue for a clearer definition of anonymity that emphasizes the aspects of online identities presented as much as those that are hidden, both for individuals presenting themselves as anonymous and for those with which they interact. This account, it is argued, not only matches what SIDE and online disinhibition expect of anonymity, but allows for a more robust empirical agenda in the study of anonymity in online political talk. It provides little support to the trade-off perspective on anonymity.

The Problems with “Trade-off” Anonymity

In cases when the “trade-off” conceptualization of anonymity is defined, the most widely cited definition is Anonymous' (1998) which describes anonymity as a single continuum ranging from fully anonymous to fully identified. In practice, anonymity is often treated as a simple binary variable. For instance, Leshed (2009) looked at policy changes between one-use pseudonymous and real name conditions, treating each condition as a natural proxy for the entirety of anonymity. Williams (2005), in an otherwise thorough theoretical examination of the influence and importance of online

anonymity (deindividuating effects as well as protecting valuable personal information) neglects to directly define anonymity. Bernstein et al. (2011) construct an implicit continuum of anonymity in their analysis of 4Chan, where websites with named users are less anonymous than pseudonymous communities which are less anonymous than 4Chan, a website where most content is presented entirely without pseudonyms. Similarly, Cho & Acquisti (2017) use real names, pseudonyms connected to social media profiles and pseudonyms unconnected to external locations as sources of increasing anonymity under the assumption that social media accounts will be intrinsically more identifiable than accounts linked only to a single website. Where anonymity is considered as a non-binary phenomenon it is often presumed to proceed along a single continuum from fully anonymous to real name conditions in line with Anonymous' (1998) definition.

Clark-Gordon et al., (2019) conduct a systematic review of the relationship between self-disclosure and anonymity in online blogs. The authors are careful to distinguish between multiple types of anonymity and, citing Anonymous (1998), frame anonymity as a continuum. This distinction, they note, is in opposition to the majority of the studies they examine, which treat anonymity as a simple binary. However, following literature used in the review, the distinction between anonymity and self-disclosure itself remains problematically unproblematized. Revealing certain information about oneself (e.g., name, social status, appearance) is a matter of nymity and revealing other information (anecdotes, personal opinions) is a matter of self-disclosure without clear conceptual distinctions between the two activities.

Rainie et al. (2017) canvassed over 1500 academics and industry professionals

about the future of incivility online. The authors include significant discussion of anonymity in their analysis of the interviews, but do not reconcile an apparent paradox that experts seem to believe is at hand: that anonymity both no longer exists and is ubiquitous. A more careful delineation might note that the vanishing anonymity is largely a back-end function of companies attaining ever greater access to personal data, while often allowing users to retain the appearance of visible anonymity in public-facing discussion. These are, perhaps, not the same anonymity at all, but rather, the use of a lay definition in much more technical circumstances, thereby blurring mixed conceptual and operational components.

Following the above literature then, a somewhat uncharitable model of “trade-off” anonymity presents itself. Anonymity is ostensibly a continuum, but practically speaking a binary, between named and anonymous conditions. When technology allows the use of pseudonyms or hiding names, users are anonymous; when it mandates legal names, users are known. In this manner, anonymity is technologically determined. Anonymity itself may be measured principally through a single convenient information channel that may vary with the researchers’ goals (e.g., Leshed, 2009; Cho & Acquisti, 2017; Bernstein et al., 2011). Meanwhile any other revelation of personal information to the same audience may be called self-disclosure or visibility or some other unclearly differentiated concept. Yet, these also point to being known in an online community, perhaps in more meaningful or revealing ways than revealing a name. This is generally a user-facing feature, however, in technical domains, anonymity can be understood alternatively purely by corporate access to data without reference to public facing data. While few to no

studies fall into every single pitfall, each issue is quite common. Needless to say, this is a somewhat unideal situation as a foundation for future research. That is to say that anonymity is in great need of a more careful concept explication (Chaffee, 1991).

To begin this process of investigating anonymity, two clear strategies exist. Firstly, anonymity can be explicated more effectively through careful delineation of different types and components of anonymity. To do this, it is necessary to review exactly what anonymity is. This will be accomplished first by investigating formal definitions of anonymity in the literature. Notably, while work defining anonymity in more careful terms exists, definitions have not generally attempted to engage with all of the issues outlined above (e.g., Anonymous, 1998, Yun, 2009, Asenbaum, 2018). Moreover, even when conceptual definitions have sought to be moderately more robust, anonymity's treatment in empirical work has often failed to actualize appropriate details. Secondly, and just as importantly, any reexamination of the construct must also consider the causal linkages proposed between anonymity and its theorized effects (Chaffee, 1991). In the literature on anonymity and online deliberation, several theoretical groundings are offered for why anonymity may lead to both good and bad outcomes. Two of the most prominent perspectives are deindividuation and the social identity model of deindividuation effects (Postmes et al., 1998) and the online disinhibition effect (Suler, 2004). Both will be evaluated in turn, with causal mechanisms highlighting shared requirements for a useful definition of anonymity. Finally, recent work identifying anonymity as a type of performance (Asenbaum, 2018) will be used to help construct a novel definition of anonymity.

Past Attempts to Conceptually Define Anonymity

Among the most widely cited definitions of anonymity used in computer-mediated-communication (CMC) research stems from Anonymous (1998). Here, anonymity is defined as “the degree to which a communicator perceives the message source is unknown and unspecified” (p387). This definition places anonymity as a purely perceptual characteristic; rather than a technological one or a matter of factual knowledge and also emphasizes that anonymity exists along a continuum from fully identified to fully anonymized. Anonymous (1998) distinguishes between *self-anonymity*, the perception that someone is anonymous, and *other-anonymity*, the perception that others are anonymous. Additionally, either self- or other-anonymity can be *physical* or *discursive*, with the former relating to the perception that one is visible, their appearance is known, or they are seen by others, and the latter concerning the extent to which one perceives their messages to have an identifiable source. Notably, this definition presumes a singular “real” identity that individuals can be linked with, devoting great attention to the presumed effect of its absence without effectively theorizing “real identity” in any substantive sense. While this definition is often cited as the justification for “trade-off” anonymity, it does not entirely match with how trade-off anonymity is most commonly operationalized.

Yun (2006) improves on this model in two ways. Firstly, and straightforwardly, Yun offers a contrast between *technical anonymity* and *perceived anonymity*. The distinction between technical systems that actually obscure identifying information (e.g., a lack of IP addresses, the presence or absence of email verification, whether one’s real

name is actually known to other members of a community) and the perception that this information has been obscured (the belief that one is unknown either by their identifying information or more generally as a recognizable member within a community) allows for a clear delineation that acknowledges independent physical and psychological conceptualizations of anonymity. Through this distinction, Yun finds that technical anonymity is a prerequisite, but not sufficient, to create perceived anonymity, and that it is perceived anonymity which directly influences individual behavior.

Yun (2006) also divides Anonymous' notion of discursive anonymity further into distinct concepts of self-anonymity — whether one's offline biographical information is known to a community — and discursive anonymity — whether one's personality and writing habits are recognizable. Yun argues that perceived anonymity does not always strictly follow from technical anonymity, and that perceived anonymity has a more direct influence on individual behavior. By highlighting conditions where, for example, self-anonymity may be high but discursive anonymity is known, Yun's framework allows for situations where an individual feels recognized and known within an online community while still hiding their offline names.

Other, broadly similar conceptualizations of anonymity exist. Marx (1999) follows Anonymous in adopting a conceptualization of anonymity as a continuum from fully anonymous to fully identified, however, Marx opts to delineate seven specific categories of identity information, that is, information by which one can be identified. Of these, many markers point specifically to a single identity or person; though distinctive behavior, social categorization and knowledge of certain symbols serve, instead, to

identify individuals within groups. Marx does not distinguish between technical and perceived anonymity or attempt to fit these categories of identity information into a more coherent conceptual framework.

Examining the practice of creating one-use “throwaway” accounts on Reddit, scholars (e.g., Ammari et al., 2019; Leavitt, 2015; Pavalanathan & Choudhury, 2015) find that users feel relatively identified on long-used pseudonymous accounts and turn to more anonymous ‘throwaway’ accounts to discuss sensitive issues, believing primary pseudonyms are not perceived as safely unlinked from offline identity and wishing to protect their primary pseudonyms’ reputations. That is, while all accounts on the platform Reddit ostensibly have technical-anonymity in Yun’s revisions to Anonymous’ (1998) framework, the notion of discursive anonymity explains why individuals may choose to disclose only in comparatively short-lived accounts. Likewise, Yun’s conceptual distinction between technical and perceptual anonymity clearly matters in such cases. Technical anonymity is unchanged from account to account which exist within the same policy regime and platform features; however, perceived anonymity can vary greatly within the same technological system. That is, someone may have a durable commonly used account that they post routinely on and feel known on, within several communities. They may also have an ephemeral throwaway account they use just once, confident that it cannot be linked back to any larger portrait of their real identity. Naive anonymity research methods that presume the affordance of anonymity entails its perception necessarily would miss the vital distinction between account types within the same platform.

The preceding definitions have largely offered compatible, if unique, conceptualizations of anonymity in terms of what information is disclosed either technically or perceptually, as evidenced by research into different account types within the same platform. Together these definitions offer a more robust conceptualization of anonymity than anticipated by binary measures or the trade-off school of thinking. They ultimately frame anonymity in terms of the degree of information that one believes hidden or known. This implicitly — if not necessarily intentionally — frames anonymity as a matter of connection not just to an identity, but to a particularly holistic notion of identity that privileges the “true” (that is, offline) self. There is no a priori reason to think this must be the case. Someone may very easily present aspects of their personality that they hide in everyday life from coworkers or even friends and family when acting online. For instance, a transgender person who tries out a new name online is anonymous in the sense that they are not using their “real” name but may feel far more visible and known than they do when using a deadname “IRL”.

Moore (2018) proposes a dimensional approach to anonymity that challenges this presupposition. To do this, Moore divides anonymity into *traceability*, the ability to connect one’s actions to their real identity; *durability*, the ease by which new identities can be changed or adopted; and *connectedness*, the extent to which identities are localized to a specific environment. Of these, identity information linking back to offline identity can be entirely found within traceability as a component of anonymity. Connectedness refers to the discreteness of individual identities, drawing back to the phenomenon of context collapse (Marwick & boyd, 2011) and the ability to maintain

multiple distinct selves in different online and offline spaces, while durability references, among other things, the difficulty of changing or modifying an identity within a space. To Moore, anonymity stretches across both knowledge of a ‘real’ offline identity and the grounding of pseudonymous identities within online communities.

Notably, Moore believes that many of the negative behaviors associated with anonymity, particularly regarding consequences and censorship, can be linked primarily to a lack of durability which creates social accountability in communities. Moore argues that the positive effects of anonymity exist primarily in allowing netizens to limit their traceability and the connectedness of their identities. To support this perspective Moore et al. (2020) distinguish between periods of policies allowing durable pseudonyms, non-durable pseudonyms, and real name usage in an analysis of comments on the Guardian’s website and find that durable pseudonyms evince more cognitive complexity than both non-durable pseudonyms and (traceable and connected) real name posts, suggesting that negative outcomes of anonymity on the Guardian web page were predicated on cases of ephemeral, rather than durable anonymity.

Rather notably, the preceding definitions of anonymity offered by Anonymous and Marx fit neatly within Moore’s concept of traceability. By contrast, Yun’s conceptualization of discursive anonymity, as a feeling of being known within communities, while ostensibly framed in terms of knowledge about a true self, may tie more strongly to Moore’s notion of durability. At the same time, Moore does not distinguish between the difficulty in creating accounts and the actual commitment and feelings of being known one experiences within a community, in contrast to Yun who

emphasizes the perceptual dimensions of anonymity.

The above theorizing presents a wider array of types of anonymity and offers dramatic improvements over the trade-off model of anonymity in allowing for the conceptualization of many cases where certain types of anonymity exist and others do not. However, as Asenbaum (2018) argues, formal definitions have tended to conceptualize anonymity in terms of the nature and amount of information that is not presented; that is, anonymity is understood as the *absence* of certain information about individuals, or individuals' perception that such an absence exists. For most theorists, this is tied to normalizing offline identity as more important and real than online identity. Moore (2018) takes the important step of conceptualizing these absences in terms of the temporal (durable) and spatial (connected) qualities of online identity, allowing anonymity to exist in reference to pseudonyms, not merely a singular offline self. That said, while Moore's approach is a compelling advance over past definitions of anonymity, it still fails to account for the precise characteristics of online anonymous identity that, as the next section will argue, have been given incredible importance in literature theorizing anonymity in terms of its effects. An alternative approach (Asenbaum, 2018) is to instead define anonymity by what is emphasized by this selective omission of personal information. By looking at theories of anonymity's influence on social behavior, it becomes clear that the traits that remain salient under anonymous conditions are of as much, if not greater, interest than the traits hidden away.

Anonymity in Theories of Deindividuation and Disinhibition

Anonymity as Social Identity in The Social Identity Model of Deindividuation Effects

An influential theory in the development of the study of anonymity and its effects is Zimbardo's (1969) work on deindividuation. To Zimbardo, anonymity reduces self-evaluation and other-evaluation and thus lowers concern for social evaluation and reduces the influence of shame, guilt, and other such emotions on motives. Zimbardo framed this deindividuation in universal terms, appealing to a pseudo-mystical force of chaos in contrast to ordered, civilized individuation. Likewise, Zimbardo emphasized the proposed antisocial consequences of this deindividuation. This argument is largely compatible with the prior definitions of anonymity as a loss of identity. If the presumptive association between disinhibition and negative outcomes is ignored, it is potentially compatible with the trade-off perspective on anonymity more broadly. However, the theory, as outlined by Zimbardo, has had mixed to negative effects that have failed to explain the range of deindividuated behavior.

Postmes and Spears' (1998) meta-analysis of deindividuation and antinormative behavior finds that that individuals default to what would be anticipated of their social groups much more regularly and predictably than they do strictly anti-normative behavior. This, in turn, aligns much more closely with the authors' own revision of Zimbardo's theory: the Social Identity Model of Deindividuation Effects (SIDE) (Reacher et al., 1995). This theory reframes deindividuation effects in terms of social identity theory (Tajfel and Turner, 2004). Instead of treating deindividuation as a return to a universal undifferentiated chaotic antisocial impulse as Zimbardo does, SIDE

reframes deindividuation as strengthening the influence of social, rather than individual, identity on behavior. As individual behavior becomes less salient and group identity becomes more salient, behavior more closely aligns with group norms specific to one's social identity.

In terms of construing anonymity, SIDE then is essentially incompatible with a universalist “trade-off” perspective of anonymity as different group identities might facilitate differently toxic or benign behavior. Asenbaum (2018) argues that, in emphasizing the role of salient group identities in conditions of low individual salience, SIDE constructs anonymity not only in terms of the individual loss of information, but also the presence of group information left in its wake. This offers a conception of anonymity grounded in both the salience of group identifying traits and the invisibility of others. Anonymity, the theory implies, must be understood not only in terms of what is invisible, but the remainder that is emphasized by that invisibility, though SIDE notably restricts its concern over what is visible to the language of specific group identities.

Antecedents of Online Disinhibition

Often cited in literature on anonymity and its effects on online discussion, Suler (2004) proposes a general online disinhibition effect that leads to both “benign” and “toxic” disinhibition. While Suler (2004) offers a sophisticated model of anonymity as a deindividuating psychological phenomenon contrasted with other forms of identity play, much work following the online disinhibition effect has simply drawn on the perspective that the internet's affordance of anonymity leads to both benign and toxic disinhibition. Suler (2004) highlights six elements of technology that cause disinhibition, but only

defines each briefly and does not illuminate the precise structure of relationships between them.

To Suler, *dissociative anonymity* is anonymity triggered by the absence of identity cues, leading to a deindividuated self. By contrast, *dissociative imagination* is when individuals adopt highly differentiated alternative personas in online spaces by altering their self-presentation. *Minimization of status and authority* is the lack of cues to offline social status, presumed to level online talk. *Invisibility* refers to the inability to see others and also the feeling that one remains unseen by others online. *Asynchronicity* is the ability to engage when building messages free from time, sending messages without seeing the consequences immediately. Finally, *solipsistic introjection* refers to a mental state where individuals fail to perceive others online as real, owing to many of the other mechanisms described, and treat actions online as mere extensions of oneself.

While only one of these factors is labeled anonymity, both dissociative anonymity and the dissociative imagination refer to the presentation of online identity as distinct from offline, and to self-presentation arising from presenting pseudonymous selves online. Broadly, dissociative anonymity and its suppression of identity seems close to Zimbardo's work on deindividuation, while dissociative imagination — adopting highly specific traits of alternative personas based on salient identity cues — seems quite similar to SIDE. Minimization of one's own status and authority is, similarly, a form of what might be understood as discursive anonymity in the preceding literature.

In that sense, while Suler narrowly presumes anonymity has a dissociative effect that roughly parallels Zimbardo's more universalist notions of deindividuation, save for

the notable inclusion of pro-social outcomes of deindividuation, Suler contrasts a specific deindividuating conceptualization of anonymity with a myriad of other anonymities. In particular, the dissociative imagination is described as something of a direct counterpoint to anonymity. Where the former buries identity, the latter involves donning a mask and a pseudonym with distinct traits. This is, in effect, a notable break from many theorists of anonymity in that Suler specifically emphasizes individual online identity as a performance of other aspects of the self, rather than as a necessary loss of identity. In practice, much of anonymous online behavior on social media, with avatars and usernames and durable pseudonyms, may take place under conditions more similar to that which Suler predicts would lead to imagination than to anonymity. Theoretically, an interesting element of this is that Suler's conception of the dissociative imagination is quite similar to SIDE, however, while the latter perspective emphasizes salient group identities, dissociative imagination instead looks to how individual traits may be presented based on archetypal or individual features.

Notably, Suler (2001) presupposes much more individualistic effects when donning masks than the 'trade-off' conceptualization of anonymity and disinhibition his work has generally been used to support. Instead, his work more clearly aligns with treatments of anonymity that embrace a diversity of effects, though this latter claim is not immediately visible in initial descriptions of the online disinhibition effect.

Moore (2018) hypothesizes that cases of durability and low connectedness facilitate dissociation. This, Moore argues, leads to a feeling of social accountability and being known within a community, rather than being deindividuated. In this sense,

Moore's framework is, perhaps, more compatible with Suler's full set of factors than previous definitions of anonymity. Not every dissociated identity is likely to lead to both benign and toxic disinhibition. Nor can it be presumed that efforts to create technical opportunities for this dissociation would apply equally well to all possible constructions of online identity.

Similarly, three other factors in Suler's model — invisibility, solipsistic introjection, and the minimization of status and authority — all refer, to some extent, to what Anonymous (1998) termed other-anonymity: the lack of knowledge about the recipients of messages. At the same time, these factors may be better understood through comparison to concepts other than anonymity. In particular, Short et al.'s (1976) social presence theory argues that social presence — defined as the salience of other parties in mediated communication — is central to understanding the psychology of communication. While not strictly a theory of anonymity per se, social presence theory has clear overlap with other-anonymity in ways that theories focused only on self-anonymity may easily overlook. That said, Suler's (2004) theory provides relatively little argument for when exactly one factor is expected to take precedence over the other, or to underline the exact relationship between these two states. The relationship between factors concerning one's perception of their own anonymity and salient personal characteristics and the factors that relate more directly to the salience of others has not been clearly explicated.

These theories show that, far from emphasizing the stripping of all paint from a canvas, it has largely been what is added or what remains that has colored the effects of

anonymity. Nonetheless, none of these theories alone is an entirely adequate definition of an anonymity which emphasizes the nature of the mask as much as the act of hiding. What is needed is a clear formal definition of anonymity that emphasizes the role and value of the pseudonym as the default and the truly anonymous as the unusual case. In this light, the remainder of this chapter will attempt to build upon past definitions of anonymity by emphasizing anonymity as offering a space for introducing imagination and constructing identity, rather than simply hiding one's identity.

Anonymity as Identity Performance

Both SIDE and the online disinhibition effect leave significant space for anonymity, not as a pure suppression of personal information, but as a form of identity work, emphasizing, in the case of SIDE, salient group characteristics, and in the case of Suler's larger body of work, individual avatars and virtual representations with significant archetypal components that emphasize different individual aspects of personality (Suler, 2001). While the focus of SIDE is primarily themed around social identity, and Suler's disinhibition effect is profoundly individualistic in its perspective, both offer a radical transformation of the concept of anonymity from how it is conceived in most definitions.

Asenbaum (2018) attempts to reframe anonymity to a specific practice of identity performance. Instead of a question of what isn't presented, anonymity is reframed as an active process of identity creation and destruction. To Asenbaum, anonymity is a particular performance that functions to emphasize certain identities by simultaneously hiding and presenting information. This work is notable, both for the Goffmanian

emphasis on performance and masking allowing for contextual identities to be created by anonymity, and also for its emphasis on how identity is masked. Citing examples as diverse as uniforms that can subordinate individual to group identity and online discussions which allow individuals to hide or attempt to abandon their gender, age, race, etc to participate more readily as unique individuals, Asenbaum argues that anonymity can construct identity in a great number of ways that fall more in line with the presuppositions about anonymity suggested by SIDE and the online disinhibition effect than other constructions of anonymity have managed to achieve.

At the same time, while recognizing that these uses are different, Asenbaum still elects to understand anonymity through a series of dichotomies. Anonymity allows truth-telling but also deception. Anonymity allows individuality but also can be used to subordinate individual identity to group measurement. While Asenbaum treats example practices of anonymity differently, there is little to no attempt to move beyond trade-off thinking and discuss different types of anonymity, largely because the sheer diversity of situations encompassed in this theory of anonymity, ranging from demonstrations to graffiti in bathrooms to online chats to dark money in politics on to large structural processes must all fit within the singular conceptual framework. Additionally, while Asenbaum breaks from the perspective on anonymity that presumes a singular ‘true’ real life identity, he is, perhaps, too quick to assert that, when rendering individuals anonymous, “the democratic subject is temporarily relieved from the constraints of the one and only identity in the public sphere, which is subject to government surveillance and commercial targeting” (p.463).

Asenbaum notes some limitations to this claim, primarily that “the anonymity of the hood eradicating gender differences in a universalizing move enacts KKK members as default men, which deters women from participating in the Klan (p464).” However, he ultimately endorses the overall effect of anonymity to be one which removes the constraints of identity. Researchers of the internet, by contrast, have hypothesized that the internet as a whole may trend default male in much the same way (Nowak, 2018). In this way, anonymity is something that can obscure traits such as gender but may, in practice, fail to achieve the utopian promise Asenbaum argues far more often than Asenbaum admits.

Lastly, in emphasizing performance over perception, Asenbaum seems to endorse a technical perspective on anonymity over a perceptual one without fully realizing this distinction. That is, the act of performing anonymity creates anonymity to Asenbaum, not the perception that one is engaging in such a performance. Grounding anonymity in these context-dependent performances also stresses a relationship between individuals and presumed audiences which seems to preclude any possibility of the more solipsistic notions of anonymity Suler considers as falling under the realm of online disinhibition. Asenbaum sees anonymity relationally, existing between individuals and the audiences that perceive aspects of their performances. By contrast, Suler’s (2004) conceptualization allows for anonymity while alone or, specifically, through perceiving oneself to be alone; this, ultimately, contradicts the notion that anonymity requires an immediate audience.

Redefining Perceived Computer-Mediated Pseudonymous Anonymity

This section outlines a preliminary redefinition of online anonymity, reframing identity cues as core to the notion of anonymity instead of their omission. This reconceptualization is grounded firmly in the presupposition that anonymity must include identity creation in addition to destruction. It also references aspects of older definitions that serve to more clearly delimit this definition as a particular anonymity from the range of possible anonymities writ large.

Perceived Anonymity

While technical anonymity is of tremendous importance to the study of CMC, the fundamentally psychological perspectives of both SIDE and the online disinhibition effect suggest that it is *perceptions* of being anonymous which drive the majority of anonymity's effects, a presupposition supported by Yun (2006)'s survey work contrasting the effects of perceived and technical anonymity. A counterargument, premised around material safety under authoritarian structures might argue that technical anonymity can create important space for anonymous communication. However, the kinds of everyday anonymity principally discussed here have little in common with those more extreme circumstances. Ultimately, an attempt to reconcile the various, often ignored, dimensions of technical anonymity, much less to fit the range of technical anonymities to that of the perceptual anonymities discussed here, falls outside the scope of the present work. Instead, the present effort focuses on more carefully delineating perceptual anonymity so as to build a definition immediately useful in studying the concept from a social psychological perspective relating to perceptions and resultant group dynamics.

Self -Anonymity

This reconceptualization of perceived anonymity encompasses both self-anonymity and other-anonymity. In this case, self-anonymity is an individual's perception of their own anonymity (that is, a person's beliefs about what identity cues they are making visible in a given space) whereas other-anonymity is an individual's perceptions of others' anonymity (that is, what identity cues they have about others.) Given the focus on defining both concepts from a single individuals' perspective, self-anonymity and other-anonymity cannot be conceived of as mirror images.

Of these, self-anonymity is redefined on terms following Asenbaum's (2018) emphasis on both identity creation and identity destruction, as well as Suler's notion of dissociative imagination and SIDE's presupposition that group identity characteristics can be more salient in cases where hiding information suppresses identity characteristics. A shared trait across these theories, though the specific term is not always used, is that of salience. Salience refers to a thing's prominence and noticeability. SIDE treats salience as a core matter of determining what group identities may be triggered. Suler (2001)'s invocation of archetypes and avatars speaks to salient traits. The present definition opts instead to describe what is salient using the encompassing term "identity cue" to avoid circumscribing what aspects of identity can and cannot be made salient. This language has the added benefit of allowing the evaluation of different identity cues. Those more visible, such as a full 3D embodied avatar, may be anticipated to be more influential than those made less visible but still present (e.g., the same avatar as a stationary image next to a name.) Similarly, multiple consistent cues may work together to enhance the salience

of specific identity traits in this aspect. Durability and connectedness, from Moore's (2018) framework are not irrelevant here but can merely be construed as two particular forms of identity cue, with durability in particular driven not by the technical affordance of durable accounts but the perception that one is an established member of a community with a valued identity.

This focus on presentation in its own terms — rather than with respect to offline identity — facilitates a system where identity online need not defer to a presumed 'real' offline identity. In particular, this concept of identity draws from Russell (2020)'s emphasis on the way that online spaces can constitute an environment to experiment with different gender and sexuality identifications, as well as Haimson et al. (2021)'s notion of "trans technologies" as those where anonymity enables "realness" in ways that many cannot find in their offline lives while also allowing for fluid, rapidly evolving presentations of identity. No presupposition is made that identity characteristics must, in some way, defer in legitimacy to an individual's presumptive offline identity.

Other-Anonymity

Drawing from Suler's (2004) notion of solipsistic introjection, as well as Short et al.'s (1974) social presence theory, other-anonymity involves not only the cognitive and affective knowledge individuals have about others, but also the salience of others in the first place. That is, other-anonymity involves both the specific arrangement of identity cues present regarding others as well as the salience of others in a communication space.

Social presence was initially defined by Short et al. as "the degree of salience of the other person in the interaction and the consequent salience of the interpersonal

relationships (p. 65).” Though, as Cummings and Wertz (in press) note, social presence has come to encompass a wide range of related constructs. Reexamining the concept in light of more recent research, Cummings and Wertz offer a revised definition of social presence: “the perceptual salience of another social actor (p.tbd).” This combines Short et al.’s initial emphasis on perceptual salience with the additional perception that a detected other is perceived as a social other, rather than an unfeeling object or an emotionless robot. This matches with the definition of self-anonymity presented above, through the reference to salience as a defining characteristic of anonymity. It also maps to Suler’s work on solipsistic introjection and the perception that others, while potentially salient, are not always perceived as distinct social actors with their own thoughts and feelings. Perceptual salience itself shares many traits with Suler’s (2004) notion of invisibility, particularly the aspects that pertain to not seeing others, rather than perceiving oneself to be unseen. Here, the salience of interactants is presupposed to be epistemically prior to their unique traits. When others are not salient, their unique traits, whatever objective knowledge an individual has, are unlikely to matter in determining social interactions. By contrast, when other social actors are perceptually salient, they are perceived as richer and more varied individuals.

Altogether then, computer-mediated anonymity as construed here is individual perceptions of their own and others’ self-representation in computer-mediated contexts. This construction of anonymity and identity emphasizes both the differential effect of different self-representations, the possibility of hiding or revealing particular information in different audiences without respect to a singular ‘true’ offline identity and the

possibility of very different outcomes determined by the range of perceived self-representation features.

Summary

The study of political talk online has been driven in part by concerns about anonymity and its range of possible effects on public discourse, both good and bad. This chapter has argued that, to adequately understand anonymity's role in online discourse it is not enough to simply presume all anonymity functions the same, but rather that differing anonymities brought about by differing inputs and perceptions can achieve diverse ends. In order to identify how differing anonymities may form more complicated and better differentiated connections to particular discursive outcomes, this chapter reexamined the concept of anonymity from a critical perspective. Firstly, a trade-off perspective on anonymity was considered and rejected. In this perspective, anonymity is a singular influence leading to both benign (disclosure, speaking out against injustice, freedom to explore identity) and toxic (insulting, deceptive, aggressive, and other antisocial behavior) disinhibition. The conceptualization of anonymity as a binary force between identity and anonymity that drives both results was examined and rejected. Notably Anonymous' (1998) conceptualization of anonymity as a continuum concerning what personal information is hidden, as well as Moore's (2018) conceptualization of anonymity as evaluated along dimensions of traceability, durability and connectedness stand out as exceptions to this simplicity.

While the latter showed promise, actual theories of anonymity's effects, SIDE, and the online disinhibition effect were shown to presuppose a version of anonymity that

creates ample space not only for hiding information but also for making other information more salient. This was tied to Asenbaum's (2018) conceptualization of anonymity as a performance of simultaneous identity creation and destruction. Following this, this chapter offered a reconceptualization of online perceived anonymity. *Perceived online anonymity is the totality of identity information individuals perceive to be presented and not presented in computer-mediated contexts, including elements of both perceived self-anonymity — the range and intensity of identity cues an individual presents in online contexts — and perceived other-anonymity — perceptions of others' social presence.* The next chapter will extend this definition by showing how perceived online anonymity can be connected to existing theories of mediated self-representation and its effects which, in turn, can lead to the formation of hypotheses as to predictable ways in which different identity presentations — or, put another way, different anonymities — might lead to unique results with regards to toxic and benign online disinhibition. Specifically, the Proteus Effect (Yee & Bailenson, 2007) posits that individuals embodied in online spaces adopt stereotyped behaviors associated with the avatars they wear, engaging in behavior similar to what Suler (2004) posits will happen under the dissociative imagination. While the Proteus Effect is ostensibly a theory that applies to immersive virtual environments (IVEs), the theory can easily be understood in terms of the above efforts to reconceptualize anonymity in mediated environments more broadly defined.

Chapter 2: The Proteus Effect

The Proteus Effect grew out of the observation that individuals adopt behavior associated with avatars assigned to them in online spaces. That is, when users enter a virtual reality space, or play a video game, the computer modeled bodies they ‘wear’ influence users to adopt behavior stereotyped of those bodies. For example, individuals assigned more conventionally attractive avatars adopted more-confident behavior within virtual environments and engaged in more self-disclosure in follow-up interviews than those assigned unattractive avatars; those assigned taller avatars negotiated more aggressively than those assigned shorter avatars (Yee & Bailenson, 2007). Those assigned sexualized avatars reported more self-objectification and acceptance of rape myths than those assigned non-sexualized avatars (Fox et al., 2013). Those assigned child avatars reported objects as larger and identified more with childish traits than those assigned adult avatars that were shrunk proportionally to the size of children (Banakou et al., 2013). Perhaps most relevant to the issue of anonymity as concerned in political talk is the finding that individuals assigned an ‘evil’ avatar in dark robes endorsed more aggressive behavior and reported less group cohesion following a discussion task in a virtual environment than individuals assigned a ‘good’ avatar in lighter robes (Peña et al., 2009). A recent meta-analysis of 46 quantitative studies found that the Proteus Effect is robust and displays a consistent effect size of .22 to .26 (Ratan et al., 2020). While small by conventional standards, the authors note this is large in the context of media effects.

Conceived, in part, as a response to a perceived focus within CMC literature on issues of anonymity and authenticity, the Proteus Effect instead seeks to shift emphasis

toward questions of the effects of self-representation on individual behavior. Thus, the Proteus Effect, — which looks to how the cues individuals use to represent themselves in virtual worlds influence their own behavior — is broadly consistent with the research agenda outlined in reconceptualizing anonymity within the previous chapter. In principle, this paper argues that the Proteus Effect, instead of an effect unique to virtual environments, can be understood as a specific case of self-anonymity as defined in the previous chapter. To illustrate the strength of the approach, this section will review existing literature around the Proteus Effect before contrasting and contextualizing it within theories of anonymity as described in the prior chapter (i.e., SIDE, solipsistic introjection, social presence theory). This theoretical alignment will be used to propose an experiment aimed at showing that online pseudonymity within a type of text-based chat rooms often used in informal political talk can achieve different effects on discourse depending on the precise identity cues made salient to anonymous participants.

Causes of the Proteus Effect

As of yet, the mechanisms driving the Proteus Effect are not entirely clear (Ratan et al., 2020). Three explanations have been posited. Firstly, Yee and Bailenson (2007) argue that the Proteus Effect is driven by self-perception theory (Bem, 1972) where individuals learn about themselves by observing their own freely chosen behavior. As individuals engage with their own avatars and observe their own actions in the context of the avatars they use, they conform behaviorally to avatar characteristics. Alternatively, the Proteus Effect has been theorized merely as an application of priming, where the mere exposure to characters with such traits influences behavior in the fashion observed

(Peña et al., 2009). However, the Proteus Effect is significantly stronger when individuals control an avatar that represents themselves rather than when they simply see that avatar perform identical tasks (Yee & Bailenson, 2009). Finally, more recent work has argued that the Proteus Effect is driven both by self-perception theory and schema-activation, arguing that individuals connect their self-schema more closely to avatar-schema as they use avatars, and that related tasks become more likely to trigger these schemas together rather than individually (Ratan et al., 2020; Ratan & Dawson, 2016). That is, as one performs tasks using an avatar one's self-perception is entwined with their perception of that avatar and individuals will alter their self-perceptions to include perceptions of salient avatar traits.

Altogether then, research has found that users controlling an avatar, rather than watching one, experience a stronger Proteus Effect (Yee & Bailenson, 2009), and that users who with an avatar more similar to themselves, or who have the opportunity to customize their avatars experience a stronger Proteus Effect (Ratan & Dawson, 2016; Ratan & Sah, 2015). However, none of these theoretical accounts offer a clear argument for why the Proteus Effect would *exclusively* apply to specifically three-dimensional avatars controlled either in virtual reality or through movement mapped onto a keyboard. This omission creates room to study the possibility of examining the Proteus Effect on social media and instant messengers. Yee (2007) holds that the Proteus Effect appears only when users control an avatar and not when watching a character reflects the unique influence of 3-dimensional embodiment. While this emphasizes the importance of connecting one's avatar to oneself in the effect, it does not directly support the

assumption that complex virtual environments with 3-dimensional avatars are a precondition for the Proteus Effect. That is, though the effect is shown when one embodies an avatar rather than simply watching a stranger, the need for virtual worlds with complex high-cost modeled representations is not tested.

The Proteus Effect and SIDE

At a glance, the similarities between the Proteus Effect and SIDE are striking. Both theories anticipate that, in cases where individuals share an identity trait associated with group membership, users would adopt traits more closely associated with the social identity than their own, as their own identity becomes less visible. In fact, Yee and Bailenson (2007) draw this comparison explicitly in their initial explanation of the Proteus Effect, arguing that, while the underlying processes are similar, SIDE draws on connection to local groups, whereas the Proteus Effect is premised on individual self-representation. In this way, Yee and Bailenson suggest that SIDE's vagueness regarding the definition of group identity allows individual cues, a separate category, to be conflated with group identity cues. Yee and Bailenson further argue that the two can be pitted against each other, using the example of individuals assigned to a hostile message board but given attractive avatars. They argue that SIDE predicts, in such cases of mixed cues, that individuals would default to the group identity (shared hostility), while the Proteus Effect pushes individuals to express an individual cue (attractive people being nicer and more outgoing). Likewise, though perhaps less important to the fundamentally social practice of anonymity (Marx, 1999), Yee and Bailenson suggest that SIDE is dependent on users engaging in a group, whereas the Proteus Effect can happen even

when individuals are alone.

While compellingly straightforward, this account is not at all clear when the nature of social identity, as SIDE considers it, is examined in more detail. As noted in the prior chapter, SIDE's notion of group identity is fundamentally grounded in social identity theory. Here, Tajfel and Turner (1979) suggest that interpersonal communication may be envisioned to lie on a continuum between entirely determined by individual traits and entirely determined by perceived group traits. While the presence of large crowds could be understood to push interactions to further rely on these group traits, it is not the case that individuals acting alone become entirely unaware of gender, age, race, aggressiveness or other identity categories.

While SIDE presumes that the presence of groups is a factor which may increase the salience of group norms, social identity theory does not strictly presume members of a physical grouping would intrinsically identify with that group. Instead, it is merely a single cue to the salience of group identity. Reache et al. (1995) describe experiments where individuals were rendered anonymous in situations that contrasted shared identity versus distinctiveness within groups. Rather than subordinating to the dominant norm or forming a superordinate group, these individuals instead simply did not experience group identity as salient. The assumption that such group identities are readily malleable likely stems from the influence of the minimal group paradigm — the finding that ingroup and outgroup identities are quite easily induced in laboratory environments even with regards to completely arbitrary or meaningless characteristics such as being a low or high guesser (Diehl, 1990). The Proteus Effect presumes a similar fluidity of identity, that individuals

will grow attached to characteristics regardless of prior traits.

Research on mixed groups can complicate this further, Randal (2002) finds that members in organizations experience more salience of their genders as group membership is mixed in disproportionate amounts. Randal (2002) finds that, when women constitute a small minority, rather than a roughly even portion of a group, the unbalanced composition itself leads to higher salience of gender. Karpowitz and Mendelberg (2014) observe that groups can move toward minority — rather than majority — norms as other factors, such as decision-making rules, highlight or minimize the influence of minorities on discussion. Yee and Bailenson (2007) imply that SIDE requires women in this circumstance, rather than becoming more aware of their own gender, to instead acclimate to dominant male norms of behavior. However, it is hardly clear that SIDE, as originally proposed, aligns with these critiques.

Studies examining the Proteus Effect have tended to bear out the assumption that group compositions are more complicated than either shared representational norms triumphing over the Proteus Effect or the Proteus Effect replacing deindividuation within a group, as both phenomena may, instead, interact with each other. Van Der Heide et al. (2013) attribute weaker Proteus Effects as a consequence of having participants engage in their study alone instead of with human partners that could have notionally been influenced by the attractiveness manipulation used in the study. Lee et al. (2014) assigned individuals randomly to either male or female avatars within a group where the participant was always a gender minority (that is, a participant male avatar paired with two female confederate avatars or the reverse) and found those assigned male avatars

perform better at math. Additionally, this effect was found to be stronger when participants assigned male avatars were instructed to compete against, rather than cooperate with, confederates assigned female avatars. However, the effect of group composition itself was not tested directly.

By reexamining the Proteus Effect as well as SIDE under the superordinate language of identity cues the arbitrary distinctions between these theories cease to matter. Likewise, the similarity between the Proteus Effect and SIDE suggests that viewing the former specifically as a theory of virtual worlds may not be a theoretically grounded choice. Why should Peña et al.'s (2009) Proteus Effect finding that individuals with KKK avatars behave more aggressively than individuals with doctor avatars be held to be theoretically distinct from Johnson and Downing's (1979) finding, oft cited in deindividuation literature, that individuals dressed as KKK members behave more aggressively when anonymous than do anonymized individuals dressed as nurses? Instead, the Proteus Effect may be better understood as a theory of the adoption and salience of identity cues, fitting neatly into the theory of anonymity outlined above.

The Proteus Effect and Online Disinhibition

In laying the groundwork for the concept of dissociative imagination, Suler (1999) predicts effects very similar to the Proteus Effect, though created by static images instead of more complex avatars. Here, Suler credits not the unique influence of embodiment in a fully realized animated virtual world, but the cognitive task of adopting a body in a much more rudimentary online space. Turkle (2011) similarly notes that users often roleplay online personas in text-based environments, quite similar conceptually if

not mechanistically to the fully embodied 3-dimensional avatars studies of the Proteus Effect typically entail. Even in the context of the gaming environments which the Proteus Effect is primarily concerned with, avatars are visual assemblies that offer a wide variety of identity cues beyond appearance. They may have names, voices, histories and personalities. They are designed with distinct body language and ways of interacting with the world. In this sense, the choice to emphasize *visual* self-representation to the exclusion of all else may limit understanding of the Proteus Effect, and of the breadth of its application. Holding that the effect requires very specific arrangements of visual representation and presuming, rather than testing, that the effect will not appear outside of these contexts is likewise limiting.

Perhaps the greatest point of contention between the online disinhibition and the Proteus Effect is that Suler (2001) draws on a much wider notion of individual traits than have appeared in the Proteus Effect literature to date. Rather than mirroring offline group dynamics or known traits from social psychology literature, Suler (2001) extends the preliminary idea of dissociative imagination to include a much wider variety of nonliving and nonhuman forms. Indeed, an avatar can just as easily be a werewolf, Zeus, or the Planet Earth; as it can be an approximately attractive or unattractive humanoid. Ratan et al. (2016) do contrast more supernatural examples, using superheroes, instead of common interpersonal norms and find evidence of the Proteus Effect. However, this example is a relative exception within the Proteus Effect literature, and even then, still hews much closer to the real-world baseline than Suler's (2001) framework for thinking about the range of possible avatars. Once again, these connections favor viewing the Proteus Effect

as an extension of a superordinate concept of anonymity influencing behavior through salient identity cues rather than through specific theories of the effects of online avatars.

Modality as a Moderating Factor in the Proteus Effect

If anything, these theoretical accounts suggest that features such as modality and control schema may be better treated as moderators, influencing the strength of the effect. “Avatars” need not be treated as restrictively as the concept often is within the Proteus Effect research when it is hard to point to clear examples firmly contrasting the (computer-mediated) Proteus Effect with (less mediated) in person parallels. By recontextualizing the Proteus Effect within a larger notion of online anonymity, it is possible to expand the Proteus Effect beyond its traditional boundaries; it becomes possible to apply the concept to a much wider array of contexts than the narrow confines of three-dimensional virtual worlds and fully realized complex three-dimensional avatars as traditionally conceived. Chiefly, despite the long-term availability of platforms such as Second Life and VRChat, and the recent push by companies such as Meta to focus on virtual worlds, social media exchanges and online political discourse continue to take place predominantly within text-based communication.

Contextualizing the Proteus Effect as a form of identity presentation in, particularly, online spaces makes it clear that research into the effect must justify any assumption that modality is not only relevant but core to the process. The anonymity framework presented in the past chapter suggests that avatars — complex assemblages of identity markers including race, age, gender, voice, appearance, body language, and even history and personalities (Banks, 2018) — may be uniquely powerful with regards to self-

representation, but that their effect should not intrinsically be treated as unique and separate from other forms of presenting identity online. Rather, the effect of modality may be more naturally understood as a moderator, where ‘richer’ modalities trigger the Proteus Effect more strongly. This has been observed using meta-analytic methods to contrast studies which used virtual reality and computer desktop environments as spaces to trigger the Proteus Effect (Beyea et al., 2022) but not examined directly with less rich modalities. Under the framework of identity cues outlined in the previous chapter with respect to anonymity, a similar moderator effect can be assumed. That is, it is possible that by adding more salient identity cues or increasing the number of identity cues present which point in the same direction, the Proteus Effect could be stronger in cases of richer modalities than in ‘lean’ modalities such as text-based chat. By problematizing the Proteus Effect’s relation to particular media, it becomes possible to directly interrogate what aspects of the theory matter. Are visual representations, direct control of avatars’ movements schemes, and 3-dimensional virtual worlds where avatars can interact necessary to the effector simply the environment where the phenomenon happened to be previously observed and subsequently presumed requisite? Could similar effects to those seen in rich embodiments be achieved with much leaner media, potentially with those that include no visible component to self-representation at all?

In the exceedingly rare cases when static image representations have been tested with the Proteus Effect, studies have found mixed results. Van Der Heide et al.(2013) found weak evidence of the Proteus Effect when manipulating avatar attractiveness in online dyadic text-based communication, but raised concerns that the awareness of a

partner's appraisal and atypical measurements in the study may have led to a weaker observed effect than prior Proteus Effect studies. Beyea (2019) found that aggressive static image avatars could induce more aggressive behavior in commenting on message boards, but their attitudinal measures did not show increased aggression following from their manipulation. Finally, Ratan et al., (2016) found evidence that using an ideal self or superhero-themed avatar increased student motivation during avatar use tasks in text-based group discussions.

The dissertation will extend the study of the Proteus Effect by testing whether similar effects to those observed with visual representations in sophisticated virtual worlds can occur in much more constrained self-representations. Finding evidence of the Proteus Effect in online political discussion would serve to connect the behavioral changes brought about by self-representation to largely empirically untested notions of anonymity and dissociation previously discussed. To highlight the significance of these theories to political talk, this dissertation will use self-representations that have been shown to induce aggression and reduced group cohesion in prior Proteus Effect work. Following closely the task and dependent variables used within a similar study by Peña et al. (2009) will help to connect this work to past observations of the Proteus Effect. Manipulating features common to many social media platforms, users will be assigned an aggressive or unaggressive username and avatar. Following Peña et al. (2009), the following hypotheses are expected to hold true within text-based environments.

H1. Users assigned aggressive accounts will (a) behave more aggressively, (b) report higher endorsement of aggressive norms and (c) report less group cohesion than those assigned unaggressive accounts.

Self-anonymity presumes that increased salience of identity cues tied to an account will lead to stronger effects on self-expression. Based on the definition of anonymity theorized in Chapter 1, it is anticipated that more congruous identity cues will strengthen the salience of an account identity. Accordingly, the following additional hypotheses are made:

H2. These effects will be stronger when users are given both visual and name-based self-representations online, than when users are only represented by usernames.

Drawing on the prior discussion of causes of the Proteus Effect as well as SIDE's relation to social identity theory, it is predicted that the extent to which one identifies with the assigned account will moderate the relationship between account assignment and adoption of behavior (Ratan & Dawson, 2016), leading to the following hypothesis:

H3. The effects predicted in H1a, b and c will increase as identification with assigned accounts increases.

The Complicating Role of Other-Anonymity

One concern in experimentally manipulating the cues driving the Proteus Effect within group contexts is that changes in relationships may be observed because of the influence of adding visuals that depict others, rather than of that depict participants themselves. Or rather, while the primary goal is to experimentally manipulate the number

of cues participants have of their own representations, doing so in small-group settings will also manipulate the number of cues participants have of their conversation partners' representations. Broadly, it is difficult to vary the levels of information about both self and other while only influencing self-anonymity and not other-anonymity. This, in turn, leads to concerns relating to both the salience and identity aspects of other-anonymity as presented in the preceding chapter.

While the Proteus Effect has been traditionally concerned with what this work describes as self-anonymity, effects cannot be entirely disentangled from other-anonymity occurring at the same time. Social presence theory (Short et al., 1976) anticipates that richer environments which include visual representations of individuals will lead to more awareness of other participants as social others. Similarly, Suler (2004) anticipates that invisibility, conceived as the absence of the immediately felt presence of others, will lead to more disinhibition. By making group members more salient, it is entirely possible that individuals will be more conscious of others as distinct people which, in line with social presence theory's expectations, could reduce incivility. As the added cues further emphasize the similarity between individuals, and group members, it is possible that this could, instead lead to an increase in favorability toward group members via greater emphasis of a shared ingroup identity. Ultimately, no specific hypotheses concerning social presence were included in the preregistration of the current study. However, to eliminate the influence of other-anonymity on the manipulations, social presence must be considered and included in these models as a control variable.

Chapter 3: Methodology

This study consisted of a preregistered experiment using a 2 (avatar and username vs. username) x2 (aggressive vs. unaggressive account details) between-participants factorial design. Following Peña et al. (2009), participants were assigned to participate in an online discussion in groups of three to discuss how bad behavior on that platform should be punished. However, where Peña et al. (2009) used an online game as the venue, this study used the instant messaging platform Discord.

While less prominent than Facebook or Twitter, Discord is an instant messenger program that provides both voice and text services for large online communities with more than 140 million monthly active users (Lunden, 2020). Additionally, the platform has been widely linked to organizing political action such as Black Lives Matter protests (Griffith et al., 2021) and to organizing far-right political violence such as that of January 6th, 2021 (Peters, 2021). Discord itself is organized topically, with individuals able to create and manage their own servers, which can range in size from 1–2 members to tens of thousands of members. Servers are primarily moderated by individual server owners rather than the platform itself. The arrangement of topical features allows for a number of explicit activist communities, but also the kind of hobbyist communities that Wright and Graham (2016) identify as an important arena for the study of informal everyday political talk online.

Participants discussed how a member of an online platform who routinely mocks new users should be punished from a list of five punishments ranging from no punishment to being permanently banned from the community. This topic was chosen,

both for its similarity to a previous Proteus Effect study (Peña et al., 2009), and because, in asking about just punishments, and how/when communities should restrict members' speech to protect other members, the topic creates room for everyday political talk despite its lack of connection to formal political issues. Groups were assigned to conditions where members did or did not have avatars depicted next to usernames. Additionally, all users within a group were assigned either an aggressive username (and avatar) or an unaggressive username (and avatar), with aggressiveness determined based on the results of a pretest.

After the discussions, participants completed measures of social presence, identification, group cohesion, aggressiveness, as well as manipulation and awareness checks.

Participants

A total of 141 participants were recruited from the BU College of Communication SONA Research Participant pool in exchange for course credit. In order to secure a comparison sample distinct from the overused population of college undergraduates, an additional sample of 72 participants were recruited using Facebook advertisements targeted at adults currently residing within the United States in exchange for modest compensation (a choice between a 10% chance of winning a \$50 Amazon gift card or a guaranteed \$5 Amazon gift card.) To ensure fairness to all participants, those recruited in exchange for course credit were given the option to receive the gift card instead as desired.

A number of participants recruited through Facebook were noted to have IP

addresses outside the United States, which should have prevented seeing the recruitment ad directly. This likely occurred due to organic sharing of the Facebook advertisements within communities of those doing such studies for compensation. Notably, a substantial fraction of these participants seemed to have attempted to participate in the study multiple times as identified by shared IP addresses, registration emails and, in multiple cases, expressed familiarity with study instructions that had not yet been presented to participants. While several of these participants were successfully prevented from participating in the study, others were not, mandating the need to remove all data from participants outside the United States entirely. By contrast, only one case of a repeat participant was observed in the sample with IP addresses within the area targeted by advertisements.

Given the fact that these participants fell outside the sampling criteria, the contingent nature of the discussions, and potential influence from those completing the study multiple times, all groups containing at least one participant identified as being outside the US, as well as the other group with a repeat participant, were removed from the data set. Two other groups that completed the study were removed from the data set due to technical issues during the deliberation process rendering data unusable. This reduced the sample to 138 participants from the student research pool and 33 Facebook-recruited participants.

Participants were also eliminated individually (rather than in groups) if they correctly identified the study manipulation when asked and noted that they had been aware of the manipulation prior to a manipulation check question within the final

questionnaire. This left a final sample of 113 student participants, and 23 Facebook-recruited participants.

Student participants had a mean age of 20.95 years old, with participants between 18 and 27 years of age. In terms of gender, 99 participants were female, 12 were male, 1 was nonbinary and the last indicated they preferred not to answer the question; 62 participants identified as Asian, 43 as White, 8 as Hispanic, Latino or Spanish Origin, 4 preferred not to say, 3 identified as Middle Eastern or North African, and 3 identified as Black or African American. Participants were also asked to identify experience with similar instant messenger software: 41 participants had used Discord before, 66 had not used Discord but had used at least one type of other similar software, and 6 participants had no experience whatsoever with similar software.

By contrast, Facebook-recruited participants had a mean age of 43.35 years old, with participants ranging from 18 to 74 years of age. In terms of gender 16 participants were female and 7 were male; 12 participants identified as White, 4 as Hispanic, Latino or of Spanish origin, 3 as Asian, 3 as Black or African American, 1 as an American Indian or Alaska Native, and 1 preferred not to say. Asked about past software experience, 12 had used Discord before, 10 had used similar software and only 1 participant had used no comparable software.

Materials

Avatar and usernames designs.

Participants were assigned an avatar/username combination chosen from a larger list. Peña et al. (2009) gave users identical avatars within condition; however, this was

accomplished in a group context where avatars were primarily identified by a shared uniform. In contrast, instant messaging avatars tend to be distinct. Without distinct usernames and avatars, identifying which participant is posting which message would prove difficult in a chat program. As such, instead of using identical avatars for all participants within each experimental condition, participants were randomly assigned one from a larger set of 12 aggressive or 9 unaggressive avatar/username combinations that qualified under pretest criteria. To preserve the aggressive/unaggressive manipulation in text and avatar-based materials, each avatar was given a corresponding thematically similar username.

Sampling and Pretesting of Stimuli.

Following Reeves and Geiger (1994)'s suggestion to improve media stimuli by varying the content of stimuli and using different media images at random, researchers created an initial list of 72 avatar and username pairs using images from pop culture, stock photography and images of characters created from popular "picrew" programs designed to function like dress-up dolls. To evaluate these avatars, a pretest was distributed online to a convenience sample recruited through postings on the r/samplesize community on the social media platform Reddit.

Each pretest participant was asked to rate a randomized sample of a mixture of 25 avatar/username pairs and usernames alone on six 7-point Likert-type questions, five of which were drawn from the items used in Nowak and Rauh (2008). These questions were used to verify avatars and username pairs were both realistic and perceived as similarly aggressive while also understanding any potential differences that could have caused

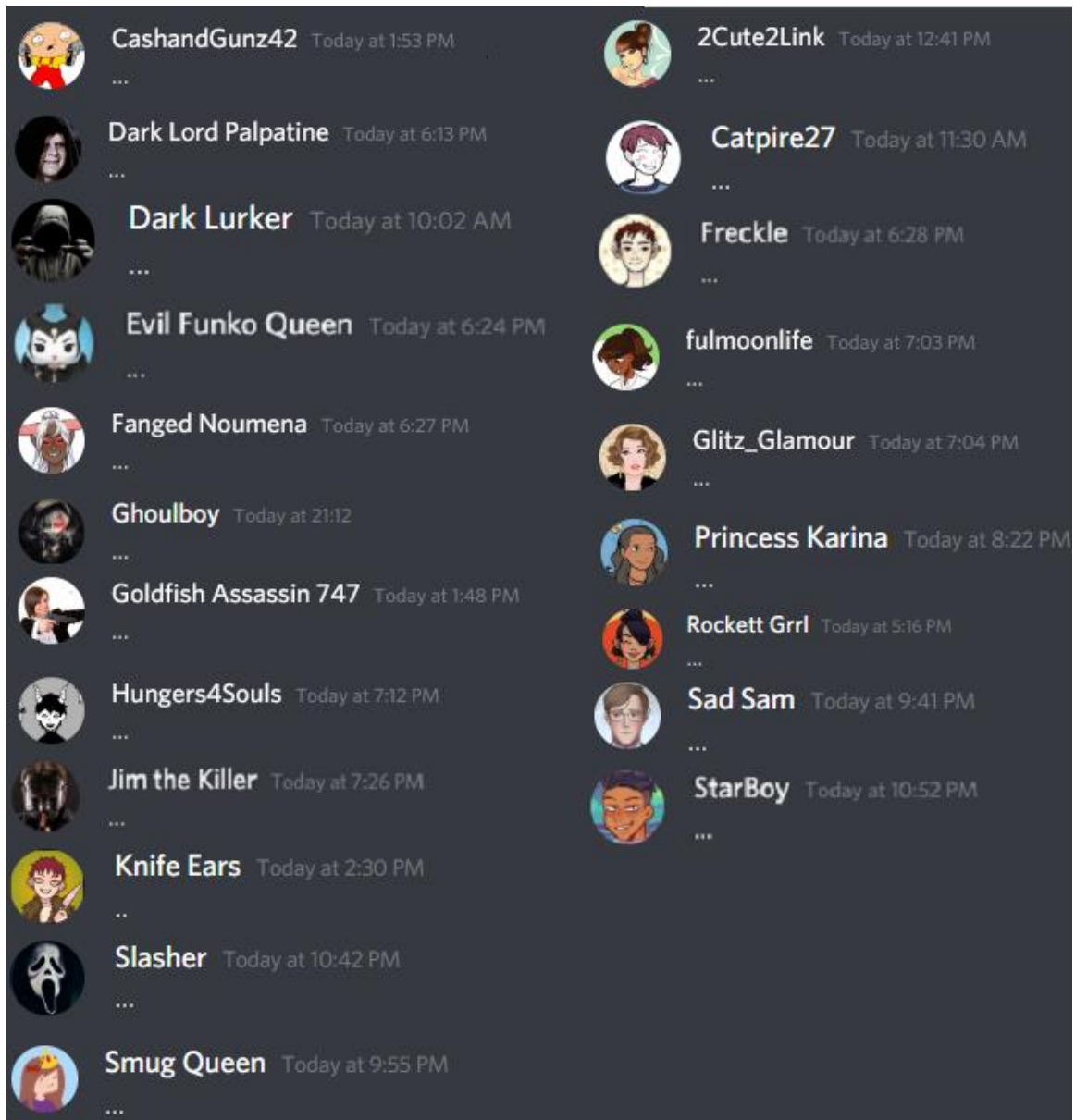
confounds in the selected manipulation. Specifically, users were asked to rate the extent to which they agreed with statements that avatars were aggressive, masculine, feminine, intelligent, reliable, and seemed like realistic avatars (the last of these was original to this study). The set of accounts used in the pretest consisted of 72 avatar/username pairs which were either presented as avatars with usernames or usernames alone required a total of 144 distinct possible objects to rate. No participant evaluated the same username both by itself and then paired with an avatar. A total of 81 participants rated avatars, though 40 participants did not successfully complete all 25 requested ratings. In the end, each username or username/avatar pair had an average of 11 ratings in total.

Avatars and username pairs with a greater mean difference than 1 scale point in terms of aggression between usernames presented with avatars and usernames presented by themselves were eliminated from the study. Any pair with a mean difference greater than one on more than 2 other dimensions was also excluded. Following this, the 12 most and least aggressive avatars were selected. Researchers electively removed another 3 avatars that were taken as differing in design from the remaining images (e.g., a realistic professional headshot of a non-famous person paired with a realistic sounding name was cut after only one of which survived the earlier elimination process), leading to the final set stimuli. For a full listing of the final set of included avatars, see Figure 3.1 on the next page.

Task. A modified version of Peña et al. (2009)'s manipulation was used in this study. Participants were initially presented with a scenario in which a member of an online community had repeatedly engaged in belittling newer members of the

community. Participants were then asked to imagine themselves as community moderators in charge of enforcing the rules. Participants were given a set of five possible punishments (no punishment, demand an apology, warn the perpetrator that they would be banned if the behavior recurred, ban them for one week, permanently ban them from the community) for the behavior described and asked to come to a consensus as to the ranked appropriateness of each for punishing the offender. Peña et al. (2009) used a similar prompt, albeit focused on violent behavior within a video game. By adapting a more general situation that encompasses participants' views on civility, punishment, and restrictions on speech within a community, the topic was political in the sense of everyday political talk — while it lacked direct commitments to electoral politics and policy, participants were asked to discuss social values in an informal everyday sense (Mansbridge, 1999). For the full script used in sessions, see [Appendix A](#).

Figure 1: Aggressive (left) and unaggressive (right) avatars and usernames



Procedure

Participants were recruited under the premise of signing up for a study of online discussion. Approximately an hour prior to beginning the study, participants were emailed a login email and password combination for the browser-based instant messaging

program Discord. Multiple sessions were run concurrently, and at least three participants were retained on a wait list each session in case those who had signed up did not log in.

Logins automatically placed participants into a prearranged chat room that could be accessed remotely on a laptop. To ensure similar experiences, participants were asked not to take part in the study on a phone. Each account had its username and avatar preconfigured to match the study conditions. Participants were asked, within the discussion room, if they had any difficulties or questions about using the software. Upon confirming that everyone had successfully logged in and could navigate the instant messenger adequately, the researcher informed participants about the scenario and stated that they should anticipate spending 15 to 20 minutes discussing. In the event that a participant did not show up, another participant at the same time slot assigned to a wait list was provided login credentials instead. In the event that no wait list participant was able to log in within 15 minutes of the scheduled start time, the session was canceled. The researcher did not participate in the conversation on their own, but monitored the conversation and answered any clarifying questions asked of them about the task or scenario. Additionally, if no participant sent a message for two or more minutes, the moderator asked if anyone had anything else to add or if the group had reached a consensus. Otherwise, after 20 minutes of discussion passed, the moderator asked if they had reached a consensus. If they had, participants were asked to give the group's rankings of each punishment. Otherwise, they were asked to report individual rankings. The moderator then gave participants a link for filling out a follow-up survey. If participants achieved a consensus more quickly, the researcher thanked them for participating and

gave the survey link. No procedure required in-person contact between researcher and participants, as the entire experiment was carried out online with participants using their own computers.

Measures

Measures consisted of both self-report posttest questionnaire items as well as analysis of the group discussion transcripts. The full self-report questionnaire is available in Appendix B.

Self-Report Measures.

Participants were first asked to identify their group number, as well as the username they had used during the study. Following this, all psychometric scales (described below) were presented in a randomized order, with individual items presented in a random order within scales. Following completion of psychometrics, participants then completed a manipulation check, followed by an awareness check, and followed by demographic questions.

Group cohesion. Group cohesion was measured using a 3-item Likert scale drawn from Seashore (1954) following Peña et al (2009)'s use of this scale. This measure captures the strength of perceived group identity based on feelings of belongingness, relationships between group members and helping (e.g., How does your group compare with other groups on the way people help each other on the task?). As in prior work, this measure was anticipated to be higher for the groups assigned unaggressive avatars. $\alpha = .8$, $M = 4.9$, $SD = 1.2$.

Attitudes toward mocking others. This was measured using a three-item scale derived from the theory of planned behavior adapted from Peña et al. (2009) that measured participants intentions, attitudes and subjective norms toward the behavior (e.g., “I would make fun of someone if I ever participated in online discussions using this account”). While these questions indirectly capture the influence of the manipulation and relate more directly to how users appraise their avatars than actual behavioral change driven by behavioral change, this measure has been established in past studies of the Proteus Effect and provides a direct point of comparison to past research. Cronbach’s alpha for this scale was marginal. $\alpha = .7$, $M = 2.2$, $SD = 1.2$.

Social presence. Social presence as a concept is intrinsically difficult to evaluate as a whole, as scales vary dramatically in how they conceptualize and measure the concept. Cummings and Wertz (in press) identify several distinct underlying conceptualizations categorized as social presence including the perceptual salience of a social actor and measures of mutual awareness and alignment. Additionally, Cummings and Wertz argue that researchers should seek to specify not only which conceptualizations of social presence they use, but how different conceptualizations do or do not relate to different outcomes. Accordingly, this study uses two distinct measures of social presence. Short et al.’ (1976) measure consists of seven different semantic differential items pertaining to the perceptual salience of social actors (e.g., “impersonal-personal,” “dead-lively”) that jointly measure both the perception of another as a social actor and salience ($\alpha = .8$, $M = 5.0$, $SD = 1.0$). By contrast, Lowden and Hostetter (2012) consists of 5 Likert-type questions predominantly concerning feelings of comfort with

other discussants, capturing the elements of mutual awareness and alignment that are alternatively conceptualized as “social presence” (e.g., “I felt that my point of view was acknowledged by other participants in the meeting”). $\alpha = .9$, $M = 6.0$, $SD = 1.0$.

Identification. Identification was measured using scales from Downs, Bowman and Banks (2019). In an attempt to synthesize disparate views of the construct, Downs et al. (2019) describe identification as a polythetic construct, where multiple independent mechanisms can achieve the same end state. In particular, they measure 6 individual mechanisms behind identification (physical similarity, value homophily, wishful identification, perspective-taking, liking and embodiment). Of these, value homophily would necessarily confound with the independent variable in question, as endorsement of aggressive norms would, effectively, be synonymous with endorsement of an aggressive avatar’s values. Additionally, perspective-taking is focused on avatars’ independent actions, which would not strictly apply in this context, while wishful identification also asked about the “kind of person” the avatar is, likewise attributing more sophisticated personality. Physical similarity would not make sense in username-only conditions and could not be used in analyses. Scales for embodiment (6 items, $\alpha = .9$, $M = 4.4$, $SD = 1.4$) and liking (4 items, $\alpha = .80$, $M = 5.1$, $SD = 1.0$) were retained as 7-point Likert type items.

Manipulation checks. To check the success of the manipulation, participants were asked to rate the accounts they used with respect to the six metrics used in the stimulus pretest. Additionally, independent sample t-tests were used to verify that no significant differences in any manipulation check variables were observed between

accounts with visual avatars and those without.

Verifying that the manipulation was successful, aggressive accounts were rated significantly more aggressive ($M = 3.2$, $SD = 1.8$) than unaggressive accounts ($M = 2.2$, $SD = 1.2$), $t(122.9) = -3.5$, $p < .001$. While significant enough to confirm that the manipulation is notable, it remains noteworthy that, despite their content (e.g., serial killers, movie villains, monsters, people with guns) the avatars in both groups were rated as more unaggressive than aggressive. No significant differences were observed in terms of how intelligent, reliable or realistic participants rated accounts based on condition (aggressive/unaggressive). However, aggressive accounts were perceived as less feminine ($M = 3.0$, $SD = 1.7$) than unaggressive accounts ($M = 3.7$, $SD = 2.0$), $t(134) = 2.15$, $p < .05$. Aggressive accounts ($M = 3.7$, $SD = 1.8$) were also perceived as more masculine than unaggressive accounts ($M = 3.09$, $SD = 1.8$), $t(134) = -1.98$, $p < .05$. This is not particularly surprising as aggression is often stereotyped as a masculine trait, and more aggressive avatars could simply have been coded as more masculine and less feminine (Williams & Best, 1990; Rosenkrantz et al., 1968).

Awareness checks. To ensure participants were unaware of the experimental manipulation, participants were given open ended questions asking them to describe the experimental manipulation as well as construct the researcher's hypotheses. A follow-up question asked them when they came to this belief during the study. A substantial number of participants indicated correctly that the avatars were involved with the manipulation, with 5 participants correctly identifying the aggression manipulation, and another 22 correctly identified that the avatars or usernames were manipulated, though could not

correctly identify why. Notably, the awareness checks immediately followed the manipulation check questions, which were likely to inform users of the manipulation. Unfortunately, while most of these participants indicated they had identified the manipulation during the survey, some explicitly identified the beginning of the survey as the point where this happened, and most did not specify when during the survey. To safeguard the integrity of data, all participants who identified the manipulation prior to the final manipulation-check survey questions or who did not specifically list the location in the survey where they identified the manipulation were removed from the data set as described in the sampling section above.

Demographics. In addition, questions regarding age, gender, ethnicity and previous experience with the technology were asked to better understand the sample's characteristics and control for possible biases.

Behavioral Measures

In addition to the self-report questionnaire, chat-logs were also coded with both a quantitative content analysis scheme focusing on the extent to which participants made supportive and disagreeing comments, and the frequency of elaboration; as well as analyzed qualitatively to identify the type of each elaboration and the broad response it received.

Quantitative content analysis was used to triangulate and supplement self-report measures. Stromer-Galley (2007) provides a versatile coding scheme for deliberation adapted to both online and offline environments. While it does not purport to directly measure aggression behavior, it is noteworthy that “aggressive” verbal behavior

measured directly (e.g., the use of swears or insults) is comparatively unlikely to appear in the context of a moderated experimental task. Instead, the instrument emphasizes issues such as whether comments are positive or negative and proportional breakdown of speaking time focusing on the valence of interactions.

In particular, previous work has found that tracking the valence of comments on others' posts can both reflect whose voice carries influence in deliberation (Mendelberg et al., 2014) as well as offer insight as to whether the overall climate of a deliberation is more or less supportive (Karpowitz & Mendelberg, 2014). The number of unsupportive comments offers an indirect measure of aggression more likely to capture the actual differences in behavior in environments where actual visible aggression is unlikely to occur. To that end, quantitative analysis focused primarily on the use of codes marking individual thoughts as agreement or disagreement.

Intended for scenarios where individuals would have voiced speaking turns of multiple minutes of length, this scale identifies turns as an individual's entire duration of speaking and codes turns as well as shorter individual thoughts that are identified by coders as statements which express a single idea within a longer turn. However, this could not be applied without modification to this system as posts were generally a single sentence or shorter and separating multiple sequential posts did not always reflect on the same speech. For example, Stromer-Galley distinguishes between turns that start a new topic and those that respond to other participants. Given the semi-synchronous nature of Discord, users could begin typing their own thought, post a fraction of a second after another user and follow up with another post that, rather than its own thought, is a direct

response to the other user. Similarly, the use of the enter button to post a message serves as a somewhat natural indicator of pauses or shifts between thoughts in many cases and avoids the difficulty in requiring coders to decide what constitutes an individual discrete thought. This scale was adapted for use in Discord by collapsing separate turn and thought categories into a single level, with each post made on Discord consisting of a singular turn and thought.

While the coding system used contains a robust array of tools to describe arguments, only three were of particular relevance here. Behavioral outcome measures used as dependent variables consisted of the observed frequency of statements expressing agreement, disagreement, and elaboration in the text chat. Agreement and disagreement, as no particular cases of overt incivility were observed, were used as weak proxies of aggression. That is, the more participants agreed with each other harmoniously, the less aggressive participants were behaving, and the more participants disagreed with each other, the more participants were likely to be engaging in aggressive behavior. Notably, this distinction is imperfect, as apparent agreement may be driven by fear of judgment (that is, fear of aggressive response) and disagreement may reflect the contrary. To account for this, elaboration — that is, giving reasons and supporting arguments for one's positions — was also included as a dependent variable, with the assumption that patterns of increased agreement or disagreement with elaboration would reflect reasoned discussion, whereas decreasing elaboration would bolster the case that decreasing agreement or increasing disagreement reflect an underlying aggression.

To render data more comparable across participants all variables were measured

as a percentage of that participant's total contribution to the discussion. That is, the word count of all messages marked as belonging to each code were first summed for each participant. Then, these word counts were divided by the total number of words each participant typed, then that sum was divided by the total words that participant typed.

Data Analysis Plan

A preregistration containing formal hypotheses and planned analyses was made through the Open Science Foundation. In particular, as the data for this project was collected within groups who interacted with each other during the stimulus, participants cannot be assumed to be independent cases. Because of this, it was planned to use a mixed effects model, also known as a multilevel regression analysis, where individual "level 1" observations are nested within a higher category of "level 2" observations and slopes are allowed to vary between groups, allowing for a better understanding of the influence of participant groups on individual behavior.

Notably, when using these models, it is important to consider mean-centering data (Hoffmann & Gavin, 1998). Individual coefficients reflect the effect of a change in a predictor value from 0 on the dependent variable. That is, the coefficient obtained in models reflects the change in a dependent variable expected following a one unit increase in the independent variable starting from 0. By mean centering these variables such that 0 is either the group or overall mean of each predictor, coefficients, instead, represent the effect of a unit change from the mean value on the dependent variable. However, two types of mean centering are widely used in multilevel models. Grand mean centering subtracts the overall sample mean from each observation, whereas group mean centering

instead subtracts the individual means within groups from each observation. Each approach has relative strengths and weaknesses. In particular, group mean centering more effectively isolates level 1 units within groups. However, Paccagnella (2006) suggests that, in cases where group sizes are small, and, particularly where researchers are interested in observing effects on individuals, rather than focusing primarily on contextual differences, it is better to avoid group-mean centering. As the manipulation here occurred at the group level — and as each group contained only a few data points meaning individual contributions to group means would be relatively large which could potentially distort analysis — grand-mean centering was used.

Handling of Multiple Samples

To best manage the inclusion of multiple sampling methods from distinct populations, it was necessary to evaluate the extent to which the two sampling methods were comparable. An independent sample t-test confirmed that respondents from the Facebook condition were significantly older ($M = 40.4$, $SD = 17.3$) than the undergraduate students ($M = 21.0$, $SD = 2.1$), $t(22.1) = 6.2$, $p < .001$. Additionally, while the majority of participants in both methods were female, a chi-square test confirmed that the gender imbalance was different across groups. 69.6% ($n = 16$) of the participants from Facebook identified as female, while this was true of 89.2% ($n = 99$) in the SONA condition. ($1, N = 136$) = 6.0, $p < .05$. However, no significant differences were noted in the percentage of participants who were white vs non-white, or in degree of familiarity with similar communication platforms. Most notably, participants from Facebook typed an average of 218.8 ($SD = 141.7$) words, while those from the undergraduate sample

typed only 179.9 ($SD = 103.2$) words on average. These differences were not statistically significant, however, that may be due to the small sample size of Facebook participants in the final data set.

Several approaches could be considered for handling this data. Most obviously, they could be analyzed separately and compared, or sampling method could be used as a “third level” grouping in the multilevel models, allowing the manipulation effects to vary between sampling method. However, the final sample of Facebook participants containing only 23 participants and the fact that there would only be two level 3 units would together render these approaches underpowered. As an alternative option, all participants were retained in the same analysis with sampling method included in the model as a control variable.

Chapter 4: Results

Per the preregistration, all hypotheses were initially tested through a series of multilevel models repeated for a) 3 behavioral variables (agreement, disagreement and elaboration), b) endorsement of aggressive norms and c) group cohesion. In each case, models were constructed iteratively. First, an empty model with no variables besides group was fitted to evaluate the percentage of each variable's data which varied at the group rather than individual level. These also served to establish a baseline of comparison for later models. Following this, variables of interest were added incrementally to the model, beginning with the independent variables. This approach allowed for granular views of the changes brought by including additional variables to the model, and construction of metrics representing the proportions of variance explained by adding these variables. Hayes (2006) uses a similar approach when illustrating the uses of multilevel modeling.

Behavioral Outcomes

Behavioral measures were used to examine aggression by observing decreases in agreement, increases in disagreement, and increases in elaboration. No evidence was found to support H1a, that aggressive accounts would lead to more aggressive behavior. Similarly, regarding H2a, that these effects would be stronger in avatar than text conditions, no interaction was observed between modality and aggression. Regarding H3a, that identification with account would intensify effects of the aggressiveness manipulation, a small, but significant effect was observed between an interaction term for perceived avatar embodiment and aggressive condition for both agreement ($p < .05$) and

disagreement ($p < .05$), suggesting that aggressive accounts may have mattered only in cases of high perceived embodiment. However, these effects were quite weak and the direction matched in both cases, contrary to hypotheses: a one unit increase in the interaction term from the mean led to a 5.8% increase in the frequency of agreement and a 4.7% increase in the frequency of disagreement. More notably, the overall suitability of these models to accurately map to the data is somewhat dubious, as Akaike Information Criterion (AIC) (which weights the inclusion of additional variables against the explanatory power gained by adding any variables) increased as any variables were added to the empty model in any of these cases. As such and given the absence of any main effect of either identification or account aggression on behavior, it is somewhat dubious to over-impugn meaning to these results. For full results concerning agreement, disagreement, and elaboration respectively, see Tables 4.1–4.3.

Results fail to reject the null hypothesis in the cases of H1a and H2a. In the case of H3a, the null hypothesis is partially rejected.

Table 1: Multilevel Models for Agreement

	Empty model	Independent variables only	With interaction term	With psychometrics	With identification interaction	With demographic variables
Aggressive		-0.04	0.00	-0.01	0.06	0.07
Avatar		0.06	0.10	0.08	-0.03	-0.03
Aggressive*Avatar			-0.08	-0.04	-0.06	-0.09
Salience				-0.04*	-0.04*	-0.04*
Association				0.02	0.02	0.02
Liking				0.02	0.02	0.02
Embodiment				0.00	-0.03	-0.03
Liking*Aggression					-0.03	-0.02
Embodiment*Aggression					0.06*	0.06*
Age						0.00
Gender (male is reference group)						-0.03
Race (all nonwhite is reference group)						-0.01
Recruitment Method						-0.01
Summary Statistics						
τ_{00}	0.00	0.00	0.00	0.00	0.00	0.00
σ^2	0.03	0.03	0.03	0.03	0.03	0.03
ICC	0.07	0.06	0.06	0.02	0.04	0.08
Level-1 pseudo R^2	NA	0.00	0.01	0.00	0.05	0.11
AIC	-52.39	-46.43	-44.24	-25.74	-20.16	-7.01

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

Table 2: Multilevel Models for Disagreement

	Empty model	Independent variables only	With interaction term	With psychometrics	With identification interaction	With demographic variables
Aggressive		0.00	0.01	0.01	0.01	0.01
Avatar		-0.03	-0.02	-0.03	0.04	0.05
Aggressive*Avatar			-0.01	-0.02	-0.03	-0.03
Salience				0.00	0.00	0.00
Association				-0.03*	-0.03*	-0.04**
Liking				-0.01	0.00	0.01
Embodiment				0.01	-0.01	-0.01
Liking*Aggression					-0.01	-0.02
Embodiment* Aggression					0.04*	0.048*
Age						0.00
Gender (male is reference group)						-0.07
Race (all nonwhite is reference group)						-0.03
Recruitment Method						-0.08
Summary Statistics						
τ_{00}	0.01	0.01	0.01	0.01	0.01	0.10
σ^2	0.01	0.01	0.01	0.01	0.01	0.01
ICC	0.44	0.45	0.45	0.51	0.53	0.90
Level-1 pseudo R^2	NA	0.00	0.00	0.10	0.16	0.09
AIC	-133.70	-124.60	-121.13	-100.25	-94.32	-69.22

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

Table 3: Multilevel Models for Elaboration

	Empty model	Independent variables only	With interaction term	With psychometrics	With identification interaction	With demographic variables
Aggressive		-0.02	-0.07	-0.06	-0.02	-0.01
Avatar		0.00	-0.05	-0.04	-0.05	-0.03
Aggressive*Avatar			0.10	0.08	0.08	0.07
Salience				0.02	0.02	0.03
Association				-0.01	-0.02	-0.04
Liking				0.01	0.00	0.01
Embodiment				-0.01	-0.01	-0.01
Liking*Aggression					0.02	0.02
Embodiment*Aggression					-0.01	0.00
Age						0.00
Gender (male is reference group)						-0.13*
Race (all nonwhite is reference group)						-0.06
Recruitment Method						0.07
Summary Statistics						
τ_{00}	0.02	0.02	0.02	0.01	0.01	0.01
σ^2	0.03	0.03	0.03	0.04	0.04	0.04
ICC	0.31	0.32	0.32	0.29	0.24	0.11
Level-1 pseudo R^2	NA	0.00	0.00	-0.01	-0.10	-0.18
AIC	-21.76	-13.39	-11.56	10.31	19.94	32.94

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

Endorsement of Aggressive Norms

Regarding H1b, that participants would report higher agreement with aggressive norms when given aggressive accounts, no main effects were observed. Similarly, regarding H2b that the aggression manipulation would interact with the presence of additional visual cues about the account, no evidence of an interaction was found. With regards to H3b, that identification with accounts would increase the effects of aggressive accounts on behavior, no evidence of interactions was observed between either identification measure and account aggression (for full models, see Table 4.4 on following page). Therefore, results fail to reject the null in the case of H1b, H2b, or H3b.

Table 4: Multilevel Models for Endorsement of Aggressive Norms

	Empty model	Independent variables only	With interaction term	With psychometrics	With identification interaction	With demographic variables
Aggressive		0.17	0.12	0.14	-0.13	-0.10
Avatar		-0.17	-0.23	-0.13	0.16	0.16
Aggressive*Avatar			0.10	-0.04	0.01	0.10
Salience				0.14	0.15	0.15
Association				-0.33*	-0.35**	-0.33*
Liking				-0.19†	-0.28*	-0.27
Embodiment				0.17*	0.24*	0.23
Liking*Aggression					0.24	0.18
Embodiment*Aggression					-0.16	-0.17
Age						0.00
Gender (male is reference group)						-0.46
Race (all nonwhite is reference group)						0.21
Recruitment Method						0.04
Summary Statistics						
τ_{00}	0.09	0.11	0.12	0.06	0.04	0.05
σ^2	1.36	1.34	1.34	1.31	1.34	1.38
ICC	0.06	0.08	0.08	0.04	0.03	0.04
Level-1 pseudo R^2	NA	0.01	0.01	0.03	0.01	-0.02
AIC	441.149	442.37	442.12	431.04	432.66	422.27

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

Group Cohesion (c)

A similar pattern emerged with regards to group cohesion. No significant relationship was observed from account aggression (H1c), and no interaction effect was observed between avatar and account aggression (H2c). As in the case of H3a concerning behavioral outcomes, a significant interaction effect of liking and account aggression was observed, lending partial support to H3a that identification with an account would increase the effects of the experimental manipulation. However, including these interaction terms in the model did not “justify” the added variables as seen by an increase in AIC when adding these variables. Likewise, with no evidence of a main effect of account aggression at any stage of the models, it remains impossible to endorse these results with high confidence (for full results, see Table 4.5 below)

Results fail to reject the null in the case of H1c or H2c and partially reject the null in the case of H3c.

Table 5: Multilevel Models for Group Cohesion

	Empty model	Independent variables only	With interaction term	With psychometrics	With identification interaction	With demographic variables
Aggressive		0.00	-0.22	-0.10	-0.16	-0.12
Avatar		0.07	-0.16	-0.07	-0.17	-0.13
Aggressive*Avatar			0.44	0.23	0.28	0.24
Salience				0.40**	0.42***	0.42***
Association				0.40**	0.37***	0.36***
Liking				0.15†	0.05	0.02
Embodiment				0.01	0.10	0.10
Liking*Aggression					0.290	0.37*
Embodiment*Aggression					-0.22	-0.23
Age						-0.01
Gender (male is reference group)						0.11
Race (all nonwhite is reference group)						-0.03
Recruitment Method						-0.35
Summary Statistics						
τ_{00}	0.21	0.23	0.22	0.05	0.08	0.10
σ^2	1.15	1.15	1.16	0.74	0.71	0.73
ICC	0.15	0.17	0.16	0.07	0.10	0.12
Level-1 pseudo R^2	NA	0.00	0.00	0.35	0.38	0.37
AIC	430.70	432.90	431.68	362.98	363.14	358.99

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

Alternative Model Specifications

The most notable issue with the above data analysis was the weak match between perceived and manipulated account aggression. That is, while participants in the aggressive group did report higher perceived account aggression ($M = 3.2$, $SD = 1.8$) than those in the unaggressive group ($M = 2.2$, $SD = 1.2$; $t(122.9) = -3.5$, $p < .001$), this difference was smaller than anticipated and neither group was rated as particularly aggressive. While it is possible that the Proteus Effect simply did not occur (or occurred so weakly as to be undetectable), it is also possible that the relative weakness of the manipulation may have masked an actual relationship. This suggests that the failure to identify meaningful relationships between the concepts under examination may have been influenced more by issues with measurement and manipulation than the lack of an underlying relationship.

O’Keefe (2003) suggests that manipulation checks, such as the measure of perceived account aggression, instead be treated as mediating variables. In most cases of media exposure, the direct effect of a technological manipulation (in this case the specific letters and pictures used to create avatars) has an effect by first influencing participant perceptions of that media, which directly influence psychological outcome variables. In light of this argument, as well as the weak effects of the manipulation, a series of alternative models were specified using the manipulation check variable (perceived account aggression) instead of manipulated account aggression.

However, a second issue arose in substituting perceived aggression for account condition within the same type of multilevel models used in the preregistered analyses. In

the case of endorsement of aggressive norms and agreement, the inter-class correlation of the empty model showed that very little variance in the data occurred at the group level (6.1% and 6.8% respectively). In both cases, the Hessian matrix failed to converge as the models, with variables specified, had no remaining variance at the group level. This indicates that multilevel modeling may have been a poor fit for the underlying data. Attempted models are provided for completeness (see Appendix C), however, interpreting either would be unreliable given the issues with group level variance. As such, hierarchical regression models are instead reported for these alternative analyses.

Behavioral Outcomes

Similar to the preregistered analyses, the alternative modeling yielded no significant main effects of perceived avatar aggression on agreement, disagreement or elaboration. However, there was a significant main effect of avatar inclusion on agreement ($\beta = 0.073$, $p < .05$) that remained when controlling for all variables. A significant interaction effect between inclusion of avatars and perceived avatar aggression was also noted when agreement was the dependent variable, though this did not remain significant once all psychometric and demographic variables were included in the model. No interactions were noted between either measure of identification and perceived avatar aggression. For full results see Table 4.6 (agreement), 4.7 (disagreement) and 4.8 (elaboration).

These results fail to reject the null hypothesis for H1a, H2a or H3a.

Table 6: Regression Results for Agreement

	Independent variables only		With interaction term		With psychometrics		With identification interaction		With demographic variables	
	B	std error	B	std error	B	std error	B	std error	B	std error
Perceived Account Aggressiveness	-0.02	0.01	0.01	0.02	0.01	0.02	0.01	0.02	0.01	0.02
Avatar	0.06	0.03	0.07*	0.03	0.07*	0.03	0.07*	0.03	0.07*	0.03
Perceived Account Aggressiveness*Avatar			-0.04*	0.02	-0.04	0.02	-0.04	0.02	-0.04	0.02
Salience					-0.04*	0.02	-0.04	0.02	-0.04	0.02
Association					0.02	0.02	0.02	0.02	0.02	0.02
Liking					0.01	0.02	0.01	0.02	0.01	0.02
Embodiment					0.00	0.01	0.00	0.01	0.00	0.02
Liking*Aggression							0.00	0.01	0.00	0.01
Embodiment*Aggression							0.00	0.01	0.00	0.01
Age									0.00	0.00
Gender (male is reference group)									-0.02	0.05
Race (all nonwhite is reference group)									-0.01	0.04
Recruitment Method									0.05	0.08
Summary Statistics										
Adjusted R ²	0.03		0.06		0.06		0.04		0.02	
ΔR ²	NA		0.03		0.00		-0.02		-0.02	

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

Table 7: Regression Results for Disagreement

	Independent variables only		With interaction term		With psychometrics		With identification interaction		With demographic variables	
	B	std error	B	std error	B	std error	B	std error	B	std error
Perceived Account Aggressiveness	0.02†	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.01
Avatar	-0.04	0.03	-0.04	0.03	-0.04	0.03	-0.03	0.03	-0.04	0.03
Perceived Account Aggressiveness*Avatar			0.00	0.02	0.01	0.02	0.01	0.02	0.01	0.02
Salience					0.00	0.02	0.00	0.02	-0.01	0.02
Association					-0.03	0.02	-0.03	0.02	-0.04*	0.02
Liking					-0.01	0.02	-0.01	0.02	0.00	0.02
Embodiment					0.00	0.01	0.00	0.01	0.00	0.01
Liking*Aggression							0.00	0.01	0.00	0.01
Embodiment*Aggression							-0.01	0.01	-0.01	0.01
Age									0.00	0.00
Gender (male is reference group)									-0.13**	0.04
Race (all nonwhite is reference group)									0.03	0.03
Recruitment Method									0.06	0.06
Summary Statistics										
Adjusted R ²	0.02		0.01		0.02		0.01		0.11	
ΔR ²	NA		-0.01		0.01		-0.01		0.10	

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

Table 8: Regression Results for Elaboration

	Independent variables only		With interaction term		With psychometrics		With identification interaction		With demographic variables	
	B	std error	B	std error	B	std error	B	std error	B	std error
Perceived Account Aggressiveness	0.00	0.01	-0.01	0.02	-0.01	0.02	-0.01	0.02	-0.02	0.02
Avatar	0.01	0.04	0.01	0.04	0.00	0.04	0.00	0.04	-0.01	0.04
Perceived Account Aggressiveness*Avatar			0.01	0.02	0.01	0.03	0.01	0.03	0.03	0.03
Salience					0.04	0.02	0.04	0.02	0.04	0.02
Association					-0.04	0.02	-0.04	0.02	-0.05*	0.02
Liking					0.00	0.02	0.00	0.03	0.01	0.03
Embodiment					-0.02	0.02	-0.02	0.02	-0.01	0.02
Liking*Aggression							0.00	0.01	0.00	0.01
Embodiment*Aggression							0.00	0.01	0.00	0.01
Age									0.00	0.00
Gender (male is reference group)									-0.16*	0.06
Race (all nonwhite is reference group)									0.08	0.04
Recruitment Method									-0.10	0.09
Summary Statistics										
Adjusted R ²	-0.02		-0.02		-0.02		-0.04		0.02	
ΔR^2	NA		-0.01		0.00		-0.02		0.06	

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

Endorsement of Aggressive Norms

The model containing only the avatar (present/absent) experimental manipulation and perceived account aggressiveness identified a significant main effect of perceived avatar aggression on participants' endorsement of aggressive norms ($\beta = .15, p < .05$) providing support for H1b, that (perceived) avatar aggression led to higher endorsement of aggressive norms. However, when an interaction term was included, this relationship ceased to be statistically significant. Instead, the interaction effect remained significant regardless of control variables ($\beta = .33, p < .05$ with all controls). This provides support for H2b, that the effect was stronger specific to the avatar condition. That said, R^2 was low overall at .073 in the model containing only both independent explanatory variables and their interaction effect. While this increased to .121 in the final model, it would be unwise to overstate the magnitude of this effect. No interaction effects were observed between either measure of identification and perceived avatar aggression, meaning that H3b was not supported. For full results see Table 4.9.

Also noteworthy is the role social presence played in these models. The perceived salience of other social actors exerted no effect on endorsement of aggressive norms, potentially implying that, even in cases where others' avatars were aggressive and salient, this exerted no influence on individuals, in contrast with the effect of their own avatars on their attitudes. By contrast, self-reported mutual awareness and alignment negatively predicted endorsement of aggressive norms. The more the group was perceived as aligned, the less participants were willing to endorse aggressive norms, despite the fact that, in cases where individuals perceived their own avatar as aggressive, the same was

likely true of their perceptions of others' avatars.

Altogether findings partially support H1b, support H2b and fail to reject the null for H3b.

Table 9: Regression Results for Endorsement of Aggressive Norms

	Independent variables only		With interaction term		With psychometrics		With identification interaction		With demographic variables	
	B	std error	B	std error	B	std error	B	std error	B	std error
Perceived Account Aggressiveness	0.152*	0.07	-0.03	0.10	-0.07	0.10	-0.05	0.10	-0.07	0.10
Avatar	-0.21	0.21	-0.22	0.21	-0.21	0.21	-0.22	0.21	-0.25	0.21
Perceived Account Aggressiveness*Avatar			0.35*	0.13	0.33*	0.13	0.31*	0.13	0.33*	0.14
Salience					0.11	0.12	0.10	0.12	0.08	0.13
Association					-0.31*	0.12	-0.31*	0.12	-0.33**	0.13
Liking					-0.15	0.12	-0.08	0.13	-0.07	0.13
Embodiment					0.15	0.08	0.15	0.09	0.16	0.09
Liking*Aggression							-0.04	0.07	-0.04	0.07
Embodiment*Aggression							0.09	0.06	0.08	0.06
Age									0.01	0.02
Gender (male is reference group)									-0.54	0.33
Race (all nonwhite is reference group)									-0.06	0.23
Recruitment Method									-0.25	0.47
Summary Statistics										
Adjusted R ²	0.03		0.07		0.12		0.13		0.12	
ΔR^2	NA		0.04		0.05		0.00		-0.01	

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

Group Cohesion

A main effect of perceived avatar aggression on group cohesion was noted ($\beta = -.19, p < .05$), supporting H1c. While the interaction between avatar inclusion and perceived avatar aggression was initially significant ($\beta = .32, p < .05$) it did not remain significant in models including psychometric and demographic variables ($\beta = .18, p < .1$), lending partial support to H2c. A significant interaction effect was observed between liking one's avatar and perceptions of avatar aggression on group cohesion ($\beta = -.13, p < .05$), though no effect was observed for embodiment. This lends partial support to H3c. As in the case of endorsement of aggressive norms, adjusted R^2 for models containing only independent variables was quite low (.063), while it was dramatically higher in the final models (.443), thus, the magnitude of these effects should likewise not be overstated. For full results see table 4.10.

These results support H1c, H2c and partially support H3c.

Adding psychometric variables caused a change of .370 in R^2 with only social presence and mutual awareness and alignment as significant, indicating that the added variables explained 37% of the variance beyond that accounted for by the independent variables. That is, as others were more salient, group cohesion increased and as the group built alignment, group cohesion increased. While the latter concept is, arguably, tied closely to group cohesion, this is not true of the perceived salience of interactants. While included primarily as control variables to isolate the individual identity component of the earlier proposed framework, the magnitude of these effects deserves notice, as does the fact that

this was a substantially distinct relationship than that between endorsement of aggressive norms and social presence.

Table 10: Regression Results for Group Cohesion

	Independent variables only		With interaction term		With psychometrics		With identification interaction		With demographic variables	
	B	std error	B	std error	B	std error	B	std error	B	std error
Perceived Account Aggressiveness	-0.15*	0.06	-0.32**	0.09	-0.19*	0.08	-0.18*	0.08	-0.19*	0.08
Avatar	0.05	0.21	0.03	0.20	0.02	0.16	0.01	0.16	0.01	0.16
Perceived Account Aggressiveness* Avatar			0.32*	0.13	0.20	0.10	0.17	0.10	0.18	0.10
Salience					0.39***	0.10	0.40***	0.09	0.39***	0.10
Association					0.38***	0.09	0.35***	0.09	0.35***	0.10
Liking					0.12	0.09	0.16	0.10	0.17	0.10
Embodiment					0.02	0.06	-0.04	0.07	-0.03	0.07
Liking*Aggression					0.02	0.06	-0.12*	0.05	-0.13*	0.05
Embodiment* Aggression							0.01	0.04	0.01	0.04
Age									-0.01	0.01
Gender (male is reference group)									-0.06	0.25
Race (all nonwhite is reference group)									0.09	0.18
Recruitment Method									0.41	0.36
Summary Statistics										
Adjusted R ²	0.03		0.07		0.44		0.45		0.44	
ΔR ²	NA		0.04		0.37		0.02		-0.01	

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

In conclusion, the alternative analyses provide a quite different picture from the preregistered analyses. As this analysis plan was not included in the preregistration it should be approached with caution. However, these results are nonetheless promising for study of the Proteus Effect in two-dimensional environments. When using the more direct measure of perceived avatar aggression instead of the experimental manipulation, significant main effects of avatar aggression are noted on endorsement of aggressive norms and group cohesion. These effects are stronger in cases where congruous identity cues are presented. However, this result should be interpreted more cautiously in the case of group cohesion as controlling for perceived levels of association between members and identification caused this relationship to weaken to the point of non-significance. By contrast, no particular influence of either perceived account aggressiveness or the inclusion of an avatar was noted on behavioral measures, leaving the results as inconsistent overall.

Chapter 5: Discussion

This study set out to examine the extent to which the Proteus Effect — the phenomenon where individuals adopt traits they associate with the avatars they embody in virtual worlds — would manifest in informally political text-based discussions on Discord. Specifically, it was hypothesized that individuals assigned aggressive avatars online would behave more aggressively in informal text-based political discussions and endorse more aggressive norms and report lower group cohesion after the fact. This study adapted methodology and measures from a prior study of the Proteus Effect that took place in an online game with more embodied avatars (Peña et al., 2009).

In doing so, this study set out to evaluate whether modality may be better considered as a moderating factor influencing salience of avatar characteristics rather than a simple precondition for the Proteus Effect to take place. To this end, the number of cues presented to participants was varied as well, with participants either shown an avatar and a username or only given a username. It was anticipated that those shown both cues would exhibit stronger effects of avatar aggression. As identification with one's avatar has been noted in the literature as a key moderator of the Proteus Effect (e.g., Ratan & Dawson, 2016), it was also anticipated that identification with one's avatar would moderate these relationships.

This study was also intended to highlight how earlier theoretical discussion could integrate the Proteus Effect into a novel construction of the concept of anonymity. This framework emphasizes that anonymity should be understood as a matter of identity cues made salient and hidden. The Proteus Effect's focus on individuals adopting traits from

the avatars they embody fits neatly into this perspective. The predictions offered by the Proteus Effect, transposed to a text-based environment, seem identical to those offered by this perspective, creating a space for theoretical consolidation within the anonymity framework proposed in Chapter 1.

Interpretation of Results

The initial registered analyses for this study broadly failed to reject the null hypothesis and no evidence of a Proteus Effect was observed with regards to any of the three dependent variables: enacted aggression (as seen through a higher willingness to disagree and less agreement), self-reported group cohesion, or self-reported endorsement of aggressive norms.

This held true across modality conditions and regardless of covariates. However, further analysis suggested that this may have been driven by the weak effects of the experimental manipulation and the poor fit of a multilevel modeling approach to the data, particularly in the case of aggressive norms. Following O’Keefe’s (2003) advice, a second set of analyses was undertaken which treated the manipulation check as the proximal cause of the hypothesized changes in outcomes, such that avatar assignments were assumed to lead to differing perceptions of an avatars’ aggressions which would, in turn, drive the Proteus Effect.

These analyses were conducted using hierarchical multiple regressions. Under these alternative analyses individuals who perceived their avatars to be more aggressive reported higher endorsement of aggressive norms and lower group cohesion, but were not observed to have behaved differently during the manipulation. Evidence that these effects

were stronger in richer modalities where visual avatars were presented in addition to usernames was likewise mixed with clear evidence of this occurring in the case of endorsement of aggressive norms, but no influence on behavioral outcomes. Results were mixed in terms of group cohesion, such that modality effects lost significance when psychometric variables were included as controls. While all reported effects were small, the presence of significant main effects offers support for the Proteus Effect.

Significant results were obtained only through self-report measures and not behavioral ones. This is likely driven by the fact that the measures used of aggression were indirect operationalizations of the underlying concept, treating agreement, disagreement and, to a lesser extent, elaboration, as proxies for aggression in the absence of clear verbal cues that could more directly be operationalized to measure aggression. Notably, Proteus Effect manipulations generally have a small effect size in richer media than that used in this study (Ratan et al., 2020). The hypotheses concerning moderation anticipated that effects would be smaller yet in leaner modalities like the instant messengers used in this study. Accordingly, it stands to reason that noise would explain the null results on behavioral outcomes. As the self-report measures were much more direct measures of their underlying constructs, it makes sense that they would be more sensitive in this case.

Implications

As these results were not obtained under the preregistered analyses, they should be interpreted with caution. That said, the findings presented in the alternate analysis section suggest that the Proteus Effect can happen in text-based environments, if perhaps

less powerfully than in the CVEs in which it has often been studied.

The Proteus Effect Occurs Through Avatar Perceptions

One inadvertent “finding” of this study is that perceptions of the experimental manipulation played a role in the Proteus Effect, while the objective experimental manipulation did not. Nonetheless, the observed results seem identical to what would be expected of the Proteus Effect. That is, these results show the Proteus Effect operates on perceptions of avatars, rather than their objective traits, though, objective assigned traits of avatars influence how those avatars are perceived. Objectively, the avatars used here were cartoonishly aggressive — featuring serial killers and menacing figures advancing toward the viewer threateningly, among others — to the point that several participants who saw only a single subset of aggressive avatars were able to correctly guess the aggression manipulation. Despite this, participants did not generally perceive those avatars as aggressive in the specific context of a chat avatar, thus blunting the effect. However, those who did perceive avatars as aggressive, regardless of their ‘objective’ aggressive traits, did respond as anticipated by the Proteus Effect.

On a pragmatic level, this calls into question elements of mediation in the Proteus Effect that have not been directly studied and suggests that the random assignment of differing avatars may be better treated as an indirect cause leading to differing avatar perceptions which then drive behavioral and attitudinal changes resulting from the Proteus Effect. Future research should deliberately engage in this kind of mediation analysis to see to what extent this influences the Proteus Effect, adopting O’Keefe (2003)’s suggestions for the study of media exposures in experimental science to the

specific context of the Proteus Effect more formally and, preferably, in advance of data analysis.

Secondly, this suggests that the role of media in the Proteus Effect may operate more as hypothesized in the earlier discussion of anonymity, that is that while technological factors and avatar assignment are important, what ultimately influences psychological outcomes are perceptions surrounding those technological factors. Or rather, it is not the objective imposition of traits that influences the Proteus Effect, but the perception that those traits have been imposed. While the two are related, perceived anonymity should be understood as the proximal cause of anonymity effects. This suggests that the Proteus Effect more clearly lines up with the articulation of anonymity in Chapter 1. It is not the objective assignment of avatar traits that matters, but the perception that those traits are present that influences how individuals perceive their own identities. This argument fits with all theoretical justifications for the Proteus Effect. Whether priming, self-perception theory or, especially, schema-activation, all accounts are driven primarily by perceived traits of the avatar, not objective characteristics. The Proteus Effect remains a specific highly studied instantiation of this broader anonymity, but it should be understood within a larger causal framework rather than through any narrowly technologically deterministic lens.

The Proteus Effect is not Dependent on a Particular Medium

These analyses suggest that the Proteus Effect can occur in text-based environments. This aligns with Beyea (2019)'s Proteus Effect findings, though that study also had mixed results. By contrast, the clearer results in favor of a Proteus Effect

occurring in text found by Ratan et al. (2016) could be attributed to the longitudinal nature of that research supplementing the weaker effects of a modality which offers less avatar cues. In doing so, it becomes harder to differentiate the effects Suler (2001) observes of avatars in online-text based environments and the Proteus Effect in richer modalities. That is, the effects of emphasizing certain personality traits via adopting self-representations that embody archetypal traits in text-based media seem quite similar to the effects of incorporating an avatar-schema into an individual's self-schema in more fully realized virtual worlds. On the other hand, the finding that the perceived salience of other social actors did not relate to endorsement of aggressive norms, and that mutual awareness and alignment instead led to a decrease in endorsement of aggressive norms suggests that the contrasting hypotheses offered by the Proteus Effect and the SIDE model cannot so easily be reconciled. That is, the study provides evidence that individual characteristics influenced behavior as hypothesized, but the salience of group characteristics did not lead to the effects that SIDE would hypothesize.

In short, these findings suggest a shift from a medium-specific approach to studying the Proteus Effect to one which examines the underlying psychological mechanisms as influenced by technology. The distinctions posited between self and other anonymity should be a key area of this approach given the initial results with regards to social presence as perceived salience of other social actors and mutual awareness and alignment. This fits neatly into the trajectory toward the study of online anonymity outlined in the earlier sections of this dissertation. By removing the technological limitations on the study of the Proteus Effect, the theory becomes, in essence, a theory of

self-presentation and its relationship with the construction of identity, as this paper has suggested anonymity must become.

Technology Moderates the Proteus Effect

The results also suggest that modality moderates the Proteus Effect. This is to say that the inclusion of images led to a stronger Proteus Effect. The effect on aggressive norms did not occur at all in text-based environments, while there was weak evidence of moderation in the case of group cohesion. These findings suggest that the Proteus Effect has been observed primarily in richer virtual worlds because the higher salience of avatar manipulations in these environments may more easily trigger the Proteus Effect. At the same time, research has not sought to distinguish the relative effects of different rich modalities as of yet, though recent meta-analytic work (Beyea et al., 2022) has found that the Proteus Effect is stronger in virtual reality than desktop environments.

The present manipulations were not able to establish that the Proteus Effect can occur in exclusively text-based environments, as effects on endorsement of aggressive norms only occurred within the avatar conditions in this study. While the theorizing around identity in this work has focused on the notion that salience — that is, how noticeable identity cues are — is the paramount variable driving images over text alone, the present work cannot adequately rule out the hypothesis that visual cues are a prerequisite for the Proteus Effect, though there is little apparent theoretical argument as to why this would be the case.

Toxic Avatars Can Lead to Toxic Disinhibition

These findings also offer clear pragmatic guidelines that may influence policy in instant messengers. Instead of viewing pseudonymous self-presentation as necessarily disinhibited, these findings suggest that, for instance, limiting the capacity of individuals to select aggressive avatar characteristics could facilitate less aggressive behavior.

Proteus Effect manipulations have been identified for a host of traits from assertiveness to gender roles, all of which might be invoked by constraining the avatars individuals are able to embody in particular online spaces. While research has yet to extend all manifestations of the Proteus Effect to text-based environments, there is little reason to think they would not transfer as was observed with aggression in this study.

While, outside of fiction (e.g., Greaves, 2022) it is generally not possible to lock individuals in a basement and coercively assign them a new social role in order to make them adopt social traits (for an exception see Zimbardo et al. (1971)), such self-presentation manipulations can be done online, often without being immediately obvious or abhorrent. For instance, the pseudonymous social media platform Reddit allows users to create avatars by combining pieces from predetermined lists of all possible options. Users independently choose facial features, skin tones and outfits from a variety of lists containing only what the platform owners have made available. Platforms could, notionally, curate the lists of available features to induce a desired Proteus Effect, such as by burying aggressive options at the back of lists while foregrounding prosocial options, or simply removing the former entirely. In doing so, platform owners could potentially nudge (Thaler & Sunstein, 2009) users toward avatars more likely to induce desired

behavioral outcomes.

This is not to suggest that platforms should engage in such tactics or even that traits such as aggression should be nudged against. In fact, while potential effects on behavior would likely be small, there is some reason to view the possibility with caution, as such attempts to influence public discussion are not strictly desirable and could be used to diminish legitimate expressions of discontent or cause a backlash. At the same time, the possibility's existence does suggest that the outcome of pseudonymity need not be (toxic) disinhibition if (toxic) disinhibition is not built for.

Limitations and Future Research

Among this research's many caveats, the most important is that results obtained were not obtained through the preregistered analysis, but rather alternative analyses selected after the preregistered analysis failed to achieve results. While these alternative analyses were justified, any results obtained through such methods must be presented and understood with caution. Relatedly, while the experimental manipulation succeeded at creating a difference in perceived-avatar aggression between groups, this difference was smaller than may be desired and no avatar was perceived to be more than neutrally aggressive. Future research should anticipate potential difficulties in inducing strong perceptions of avatars, as even images of the villains of slasher movies, pictures of people pointing guns at the viewer and an avatar depicting the sinister Emperor Palpatine of the *Star Wars* franchise were not perceived to be particularly aggressive avatars. The weak results of the manipulation despite the "objective" inclusion of a number of aggressive identity markers boosts the argument that perceived avatar traits trump objective

manipulations in directly inducing the Proteus Effect.

Likely as a consequence of the extreme nature of included avatars, a number of participants, particularly from the student sample, saw through the manipulation to identify avatar characteristics as the manipulated dependent variable, likely as a consequence of efforts to find avatars that participants would perceive as aggressive. While these participants were removed from all analyses, the exclusion of those astute enough to notice the manipulation may have biased the sample included in analyses by leaving only the subset of individuals unlikely to notice and react against or toward these manipulations. Another limitation is that the behavioral measures preselected to serve as proxies for aggression may not have functioned as such. This is less easily addressed, as few participants behaved visibly aggressive or polite to other participants. Nonetheless, the failure to successfully triangulate results across dependent variables using different measurement strategies weakens results.

The split sampling approach used for this study encountered a number of difficulties. Chiefly, participants recruited from social media often failed to meet inclusion criteria regarding being located in the United States, leading to the removal of many participants from the study. This weakened the ability to compare across samples, though no significant differences were observed between the Facebook and undergraduate samples despite their demographic differences.

Another possible confound is that participants may have found the discussion question ‘too easy’ as it was noted while facilitating discussions that many groups seemed to arrive at conclusions very quickly with minimal debate. The situation, where

experienced hypothetical community members made fun of new members for not knowing community norms was seen as one where a variety of possible valid answers could be selected, anticipating contrasting informally political arguments in favor of community good and free expression. However, many participants noted that the topic was quite “easy” in their manipulation check responses. Additionally, participants arrived at final conclusions that were largely similar. The recommended action of warning participants that they would be banned if behavior repeated was overwhelmingly selected as the best option by 72.3% of respondents, with another 20.0% identifying it as the second-best option. Not acting was overwhelmingly viewed as the worst (75.6% of respondents) or second worst (11.4% of respondents) option.

At first, these responses were seen as an indication that the discussion task may have been one where the compromise option was simply too clearly best and inaction unjustifiable. However, analysis of deliberation transcripts challenged this supposition. Notably, it was found that a number of participants very explicitly emphasized their responsibility — as moderators — to act. The hypothetical situation did ask participants to imagine being moderators, but did so without anticipating that this would do more than prompt discussion of participants’ values. At the same time, as a volunteer content moderator on a small social media platform, I have often seen that moderators feel an instinctive need to be seen taking action on all user reports that come their way, often to the concern of administrators who wish that moderators would be more willing to quickly dismiss clearly frivolous accusations of misconduct. The fact that the observed arguments invoked the specific responsibility of a moderator to act in ways that specifically aligned

with this personal account provided the impetus for a secondary examination of the data to observe and theorize these events more formally.

The following chapter details a qualitative examination of the discussion data that was aimed at evaluating to what extent these anecdotal observations bore out, as well as ascertaining whether the kind of informal political talk this study sought to investigate still took place despite the ease of the discussion task. A follow-up experiment was then designed to see if assigning participants to behave “as moderators” had influenced their behavior in ways that could be understood almost as a second Proteus Effect independent of the nature of their account condition. Specifically, assigning them to act in the role of moderators may have served to impose an additional shared additional identity cue on participants. This may have driven them to construct their online personas not only through the obvious avatar traits, but also in terms of the assigned social role of a moderator. If evidence of this can be found, it would suggest that similar identity effects can occur through theoretically similar mechanisms, despite those mechanisms being divorced from the specific assemblage of the avatar or account that participants used.

Chapter 6: Qualitative Assessment of Participant Discussions

Over the course of data collection, I noted a surprising pattern of discussion among users. The discussion prompt asked users to envision themselves as content moderators in an online community, principally to mirror the manipulation used by Peña et al. (2009). It was anticipated in advance that the specific wording of the prompt would have little effect beyond encouraging participants to discuss how they personally believe matters of mild breaches of the social contract should be handled in online discussion. However, within the discussions, many participants seemed to spontaneously begin caring quite a bit about the fictional communities they moderated, framing their opinions not just in what they believed best, but what would best satisfy the responsibility of building a good community or what might satisfy the goals of a presumed employer (e.g., attracting more members to the platform.) One participant noted that “at least as a mod, we should do something about it” to dismiss inaction even when others suggested the behavior itself might not need to be dealt with at all. Rather remarkably, after one early session, a participant noted spontaneously that the discussion experience had given them a newfound appreciation for the complexity of content moderation online that they had not previously had.

The experiment was intended to lure participants into an informally political discussion, offering a relatively non-politicized topic that nonetheless had room for strong ideological underpinnings concerning free speech, the collective vs individual good and the responsibility to stop or control online misbehavior. Instead, it seemed that participants often adopted the perspective of moderators as they construed it, engaging in

a form of shared perspective-taking that had everything to do with the prompt asking them to put themselves in moderators' shoes and potentially quite little to do with the specific individual account details assigned to users, or individual preferences for punishment in the abstract. Groups exhibited a strong trend to prefer the middle answers with surprising uniformity as the "fairest" punishment, determining that punishments must be handed out. While it is possible that this reflects surprising ideological uniformity in the sample, or simply that the discussion prompt had a generally socially preferable solution, initial examination of the data led to the inductive argument that this attitude could also have been driven by the need to act as moderators and, in doing so, to be seen doing something.

This possibility challenges the boundaries of the Proteus Effect, narrowly construed as unintentional adoption of another identity schema into one's own. Arguments invoking a moderator's perspective often did so in a way that appeared intentional. For instance, one participant wrote: "letting them know there is a possibility of being banned shows we as monitors on the platform are serious about keeping discord a safe community." This statement deliberately invokes the role of moderators as a group identity in order to persuade other participants, rather than unconsciously acting it out. Traditionally however, researchers have not directly equated the Proteus Effect to the more active practice of perspective-taking (Clark, 2020), though cases such as this seem to blur the line between unintentional and intentional adoption of alternative perspectives.

Why would moderators feel pressured to act or to be seen acting visibly? Firstly, it must be understood that the volunteer community managers that populate Discords, as

well as Subreddits, forums, and a variety of other smaller platforms differ dramatically from the systems Facebook, Google and Twitter put in place. As Gillespie (2018) notes, volunteer community managers often have a clear social investment in the success of a group, an established position of trust within the communities they moderate and a wealth of social ties to the community; all of which stand in sharp contrast to the vast and anonymized teams of paid content moderators on larger less clearly partitioned platforms. Much of the existing research on these moderators has focused on a platform perspective, analyzing the role they play in a larger social media ecosystem (e.g., Gillespie, 2018; Matias, 2019) or the effects of moderation policy or systems (e.g., Wright, 2006). As an exception, Wohn (2019) conducted qualitative interviews of a number of community moderators on the streaming platform Twitch. She found that a number of moderators frame themselves as “Justice Enforcers” who feel empowered and made responsible to take an active role in finding and punishing “bad guys.” Seen this way, it is likely that moderators may face some pressure to engage in social desirability, specifically moderators may feel a need to engage in what Paulhus (2002) describes as impression management driven by moralistic bias, where individuals are driven to act in ways that claim moral qualities to external audiences. In this sense, moderators may face social pressure to act, regardless of whether they think their actions will be welcomed by the community or are ultimately necessary, simply to demonstrate a satisfactory moral character as a justice enforcer: “at least as a mod, we should do something about it”.

To better ground these inductive observations, a qualitative analysis of the transcripts generated in the preceding was undertaken. This analysis focused on how

participants justified their stances, looking to differentiate more cleanly removed justifications from ones focused on the specific prompt to engage in discussion as moderators. At the same time, this allowed me to evaluate to what extent participants engaged in the kind of everyday online political talk that is considered to be of such importance (Wright & Graham, 2016; Wright, 2012).

Methodology

This thematic analysis of session conversation was undertaken to examine the ways that participants justified their arguments, through broader invocations of (political) ideology and ethics and invoking logics more specific to the perspective of moderators. In particular, this analysis expanded on instances where the “elaboration” category was marked in the prior content analysis using Stromer-Galley’s (2003) coding categories:

Elaboration can be in the form of further justification (as simple as: I’m for k–8, because I think it solves the problems we face), a definition, a reason for holding the opinion, an example, a story, a statistic, or fact, a hypothetical example, a solution to the problem, further explanation for why the problem is a problem, a definition, an analogy, a consequence to the problem or solution, a sign that something exists or does not exist, or any further attempt to say what they mean or why they have taken the position that they have (p10).

The qualitative analysis adopted a bottom-up approach to identify what specific *justifications* were invoked when participants elaborated on their stances. That is, this analysis categorized what arguments were invoked to justify orientations toward the respective punishments. All messages included in the transcripts from group discussions

that were coded as containing elaboration were then analyzed to see how this elaboration functioned. This analysis adopted a bottom-up approach, examining chat data and coding the arguments made into new categories until the coding scheme proved adequate to cover the entire breadth of relevant data rather than starting with a set of codes in advance.

Upon initially reading through all messages containing elaboration, I identified and coded five recurring justifications. These served to describe all instances where participants provided what I would call ideologically driven justification for their arguments (rather than, say, arguing based purely on the difficulty of taking certain actions or simply recounting an anecdote without clearly using it to lend weight to a specific justification). In light of the fact that arguments could be contrasted against each other, all codes were made distinct binary values that could co-occur. **Law and Order** consisted of arguments premised on the fact that the moderators should hand out punishment to enforce rules or use force to keep community members behaving acceptably, or simply that moderators must act because that is their job. **Community Good** concerned arguments about the relative harms and benefits to a community from certain actions such as fears of driving off members and hurting the platform, or capacity to attract new members. **Education/Rehabilitation** concerned the relative rehabilitative value of potential punishment as a teaching tool. **Victims versus Oppressors** consisted of statements that considered the relative harm to victims and those engaged in misbehavior, such as those hypothesizing that while leveling a punishment may hurt the person engaging in bad behavior, the needs of the victim should be treated with more

importance. **Free speech** referred to discussion of the principle of free speech. Lastly, a **Not Justification** category was used for cases of elaboration that did not clearly relate to justifying moderation policies on ideological grounds, such as arguments about pragmatic difficulties in implementing a punishment or statements exclusively referencing past experiences with different systems of enforcement that made no clear appeals to why a policy might be preferred. This code was not deployed with any other code.

Notably, this coding did not attempt to categorize the quality of any argument, or whether the arguer endorsed or rejected the justification they used. Many participants that invoked Free Speech did so with some degree of ambivalence, or even outright disavowal of their own points. As anticipated, the use of these various justifications often reflected an implicit or explicit adoption of a moderator's perspective. A thorough analysis of how they were deployed — in ways which these justifications warranted (Toulmin, 2003) their arguments from moderators about the duty of moderators — served to answer the questions posed about initial evidence of a second Proteus Effect. That is, while few of the justifications described were anticipated to map directly onto arguing as moderators or not doing so, a more in-depth qualitative analysis of the way they were deployed and their priors served to better elucidate the argument that participants positioned themselves as moderators for the purposes of the debate and that this influenced their stated beliefs.

As with the predefined content analysis in the prior study, each justification code was initially coded at the level of individual messages included in the transcript, then transformed to express the percentage of each individual participant's total word count occupied by messages marked with that code, to understand its prevalence in terms of

identifying what percent of the discussions had the traits of informal political talk.

Results

Participants spent, on average, slightly more than a quarter of their time elaborating. ($M = 26.8\%$, $SD = 22.8\%$). However, out of this elaboration, justifications proved rarer. Law and Order ($M = 6.1\%$, $SD = 11.5\%$) was the most common, followed by community good ($M = 5.21\%$, $SD = 10.4\%$), then education ($M = 4.5\%$, $SD = 10.0\%$) and victims versus oppressors ($M = 3.5\%$, $SD = 8.5\%$). Free speech was the least common justification discussed ($M = 1.8\%$, $SD = 6.4\%$). Not Justification consisted of 8.1% of participants' words on average ($SD = 11.8\%$). Note, these results should not imply that the hypothetical average participant spent 6% of their time discussing Law and Order, 5% discussing the Community Good, and so on. Rather, the median for each of these justifications was 0%, with a minority of participants spending a significant proportion of their speech discussing each.

Law and Order

At its core, the Law and Order justification pertained to appeals to the necessity of carceral systems to maintain communities. For instance, one user noted:

people just dont care how they actions hurt other people sometime , and they never be held accountable for they actions so they keep doing it . so i will definitely put some type of restrictions on an account i moderate for such rude behavior

This emphasizes a view of civility as something maintained by the actions of moderators to affirmatively stop bad actors. Often this premise was largely unstated. Instead, users

simply argued that it is the job of moderators to act strictly and harshly, “it is his/her responsibility to apologize to someone who get hurt by them [the person mocking new users], it is moderator's responsibility to stop this behavior” for instance, was deployed in response to statements that moderators may not need to stop behavior. Similarly, appeals to the rule of law were marked under this category.

Rather notably, this argument often contains an implicit deference to hypothetical community stakeholders that did not exist in the context of the discussion task, showcasing adoption of a moderator’s particularly constrained perspective and bolstering the case of a Proteus Effect occurring. For instance, one participant argues: “If we have a RULE against dissing, then we have to enforce it or bad behavior will spread. But if it’s just a NORM not to clown on new members, then some low-level dissing is fine imo.” This argument implies that, as a moderator it is not the participant’s job to set community rules, only to enforce the rules that a community sets for itself, or that are set by a community’s owner. Interestingly, this ignores the broader context of the discussion task, asking participants which procedures they believe are best used to handle these situations. This participant’s group proceeded to spend a significant amount of time discussing their (hypothetical) online community and articulating its norms, engaging in an almost roleplay-like mode of communication, offering clear indications of just the kind of identity adoption this analysis sought to find.

This kind of deference can, alternatively, be seen as an argument in favor of a diversity of rulesets across communities. As another participant summarizes: “to be fair, it's only an online community. It's like private clubs, they set their own rules who can join

and how to behave in the room.” Seen thus, this is less a refusal to engage in debate about how moderators should take action and more an articulation that one size of moderation does not fit all, such that different types of communities may want different rules. Nonetheless, the way these statements are structured reflects a surprising commitment to the identity of moderator.

Overall, this code represented, to varying extents, an endorsement in retributive justice as well as a deference to strong systems of rules. This provides evidence that these discussions can trigger the type of informal political talk that this project was attempting to study, which lacks many of the formal logical and behavioral commitments of ideal deliberation, but rather reflects an emergent appeal to and discussion of deeper underlying ideologies and political stances. It also highlights the ways in which participants seemed to adopt the perspective of a moderator, even to the extent that it hindered them from engaging directly in the actual discussion topic. While many cases were too ambiguous as to ascertain how often this occurred, this nonetheless offers evidence that several participants did adopt the perspective of a moderator when approaching the task, showing evidence of a Proteus Effect.

Community Good

Community Good deals with those arguments framed as benefiting the broader community, rather than punishing or helping any individual. These arguments were deployed most commonly in terms of the presumed need to increase community membership. As one participant notes: “Would we lose people faster if they were being banned or leaving on their own because of this one person?” In practice, this grounds the

participants' perspective as firmly aligned with the goals of growing and empowering the overall community, implicitly adopting a perspective more aligned with presumed interests of moderators and further strengthening the case for a Proteus Effect.

At its most extreme, this logic takes on an explicitly capitalist slant. I.e., “But would we rather have a bad customer? I wouldn't. If the customer is bad, they might drive off better customers, so I would rather they leave.” Notably, nothing in the prompt suggested that participants should envision moderators as having a financial interest in the platform's success, and, in practice, discord moderators are not generally employed by Discord at all, rather they volunteer their work on behalf of the communities in which they partake. This supposition of financial interest, and the fact that many participants chose to envision the platform's financial interests as intrinsically more worth defending than user's interests, reflect internalized beliefs about what moderators do and who they serve that were then reflected in how participants addressed the question of what should be done.

Other participants however, adopted a broader conception of community good. One participant (discussing both the harms to victims versus oppressors and community good) equates the idea of punishing individuals with creating a better climate for the majority: “Yea I think I have option 5 before option 1 just because why should the users being harassed feel unwelcomed. I would rather extremely punish one person for bad behavior than allow bad behavior to run rampant.” Others suggested that not punishing users might achieve the same effect, not changing the behavior but preserving the appearance of social harmony to the community's benefit:” If it doesn't require a

punishment, the person may not be aware of he/she's wrong; but it can help maintain a positive image of the community since it seems nothing happened.”

Others still adopted notions similar to the ‘roleplay’ discussed in Law and Order, where the individual instance of mocking was seen as an opportunity to identify what values and types of behavior a hypothetical community should be built on, and what in turn should be facilitated and enforced by these participants as moderators, rather than directly addressing the question of punishment:

A little off topic, but it is ironic to me that this individual is mocking people for not following community norms, when that action itself goes against the community norms we want to foster. This is definitely an opportunity to have a larger conversation about how we as moderators can better improve the norms.

This too illustrates participants putting themselves very directly into the shoes of moderators, considering not only how they would like to see these situations handled ethically, but what their individual responsibilities and interests as moderators should be to help the broader community. While distinct from the earlier economic logics that framed the moderators’ job as one of increasing membership, this nonetheless reflects, at its core, an act of roleplaying the moderators’ perspective, however participants happened to construct the notion of moderators as a group.

Education/Rehabilitation

The Education/Rehabilitation topic consisted of messages focusing on the rehabilitative value of punishment. That is, participants often assumed that some punishment could serve an educational value, allowing the perpetrators to know that their

actions were wrong: “But also, knowing that if you don't have something restorative in terms of letting them know their behavior is wrong, I feel like letting someone just go out into other communities to bully people is the wrong move.”

Notably, this topic was not always invoked with a potential rulebreaker's benefit in mind. In some cases participants cited procedural grounds, that it would be unfair to punish people without clearly illustrating why, both individually and in terms of preventing misbehavior in the broader community: “I think it would be unfair just to ban the user because they may not understand the issue for why they were banned. It also creates confusion among the community because they don't know why. Talking to the user and trying to get them to understand is better than just banning them without any warning.” In this sense, the focus on the educational value of punishment could easily overlap with appeals to the community good.

This topic is somewhat distinct from the prior two as participants did not necessarily position themselves as clearly as moderators. Discussing education implied that the participants would be the ones in a position to educate, not need remedial ethics lessons, but did not necessarily adopt a moderation logic in the same way seen by deferring to the authority of community rules or needing to act simply because they were moderators and moderators should act.

Victims versus Oppressors

Victims versus Oppressors consisted of posts contrasting the harm that actions might do to those engaging in misbehavior with the harm that inaction might do to the ones suffering this misbehavior. As one participant discussed this weighing process:

As a position of leadership in the online community, making new members feel welcome is a vital part of the job. Therefore, I think we should take these problems seriously and make it clear this kind of communication is not allowed in our community. The first two options (no punishment and apology) seem too lenient. However, we also have a duty to our current members and treating them with respect and empathy.

Rather notably, this category often overlapped with Free Speech, as participants often grounded the harm for punishing misbehavior on chilling free speech. As one participant noted: “I feel like with freedom of speech vs. someone being ridiculed, someone being ridiculed is more important since it directly harms the person.” Rather notably, in contrast to the supposition seen in Law and Order that moderators have a duty to act, this position, while often weighing the harms of misused speech more highly, adopted a more ‘neutral’ perspective toward misbehavior. As with education, this was not strictly refusing to act as moderators so much as there was simply no clear data to suggest participants did or did not position themselves as moderators when making these arguments.

Free Speech

The least commonly seen justification, Free Speech, consisted of those statements which considered the value of allowing unrestricted speech. Not all participants who invoked these arguments did so in favor of them, as participants often invoked Free Speech without much attendant argument. One noted: “is there probably also a like freedom of speech issue here if you keep removing them” in relation to the notion of

banning platform users. This statement earned no response, and the participant did not expand or change their opinions from prior stated contrary preferences after suggesting the potential issue. Indeed, participants did not seem, on the whole, particularly moved with the notion of free speech.

That said, some participants did evoke Free Speech arguments with a stance that was almost hostile to the idea of content moderation, in contrast to the earlier justifications that either positioned participants as roleplaying moderators or more neutrally toward moderation. One participant noted: “We already have a mechanism for regulating speech, and that is the American judicial system. I do not see why we need a second system.” suggesting that only judges should be able to punish individuals for incivility online and no system outside of the courts should be able to restrict behavior. However, this was something of an exception to the overarching norm of ignoring concerns over censorship or chilling free discussion entirely or weighing them very explicitly against alternative virtues as in the case of Victims versus Oppressors.

Discussion

This qualitative analysis was undertaken to examine the extent to which participants evinced another kind of identity adoption than that initially hypothesized. Specifically, initial observations of chat transcripts identified a number of cases where participants seemed to adopt moderators’ perspectives into their identity, reflecting, perhaps, a kind of second, if less strictly avatar-driven, case where behavior was altered by the identity cues salient to participants in the study. Initial examination of the data suggested that this second Proteus Effect could be associated in some cases with a

perceived need to act, rather than letting mild rudeness go unsanctioned, which may, in turn, have driven the surprising degree of uniformity observed in the way participants answered the discussion task. An ancillary goal was to examine the extent to which participants engaged in offering arguments that, rather than simply discussions of preference about the specific issue, tapped into or reflected broader political attitudes.

Results and Implications

Participants were observed to spend ‘only’ approximately a quarter of their time elaborating on positions, with much of this taken up by elaboration on feasibility grounds or offering anecdotes that did not clearly invoke the kind of political discussion at play. A portion of this elaboration did invoke larger themes as participants justified their arguments from a variety of perspectives. The relative frequency of this discussion could be interpreted pessimistically as an indication that such political talk was rare in this study. At the same time, Wright (2012) notes that interpreting results such as these dismally is often a consequence of implausible expectations setting impossible standards for deliberation. Instead, for the purposes of this study it is enough to show that these justifications were deployed and, while comprising a relative minority of discussion, were recurring and common enough to appear across discussion groups, suggesting that some degree of informal political talk did occur despite the apparent ease of the discussion task.

Five recurring justifications were observed to take place infrequently. Of these, the two most common justifications **Law and Order** and **Community Good** often invoked the responsibilities of moderators to justify particular plans of action for

punishing mild misbehavior online. This offered preliminary evidence that participants may have engaged in an unanticipated form of Proteus Effect, adopting, in addition to the hypothesized aggression of assigned avatars (or lack thereof), the group role of moderation prompted via text as a characteristic of the personas they should adopt during the discussion. It is likely somewhat controversial to label this behavior a Proteus Effect specifically, if the Proteus Effect is construed directly as stemming from avatar characteristics, as while participants were assigned to adopt moderator as an identity label, the trait itself was not obviously paired with their avatars. That is, instead of being asked to embody or roleplay a character who was a moderator, participants were simply asked to participate as themselves as moderators while also using avatars. The manipulation tasked participants to consciously add these traits to their self-schema, rather than assigning them a role and hoping that the traits were matched. At the same time, the underlying hypothesized process, where the salient identity label of moderator shared across the group influenced individuals to alter their self-perceptions and, accordingly, the way they argued, overlaps neatly with the Proteus Effect, and definitely fits within the anticipations of the anonymity framework presented in Chapter 1. Thus, it seems somewhat fair to use the label of Proteus Effect. However, calling what was observed in this analysis a Proteus Effect serves not to bolster the theory but to problematize yet more assumptions as to its boundary conditions.

Relatively few participants engaged substantially in justification at all, with the average participant spending only approximately 5% of their speaking time discussing each of these arguments, thus, while these results are promising in terms of suggesting

the effect could have occurred, they are insufficient to demonstrate the phenomenon by themselves. The less common justifications of **Education/Rehabilitation** and **Victims versus Oppressors** were more neutral toward adopting a moderator's perspective; at the very least they offered little clear example where this perspective was manifest in participants' arguments. The least common justification, **Free Speech**, was invoked in some cases, in explicit opposition to the very notion of content moderation.

Limitations and Next Steps

However, the limited secondary data analysis of the discussions cannot not adequately untangle the causal relationships between these actions or, more strictly, disentangle the observation that participants often cast themselves as moderators from the possibility that participants might simply have defaulted to similar terms even without an explicit identity prime. While it was possible to observe some participants adopting this kind of logic, it is not clear that this was a causal response to the prompt, nor that the phenomenon was widespread enough to merit serious consideration. Instead, this analysis should be treated as an exercise in theory generation, leading to hypotheses which a future project, such as the one in Chapter 7, could test.

In fact, the strong push toward the middle options and to ignore doing nothing might have been driven by some form of social desirability bias in light of the publicly visible nature of rankings to other discussion participants. While these observations lend some evidence to the supposition that a second type of identity manipulation had inadvertently been activated, they cannot not address this question with any great deal of control. Additionally, they cannot precisely identify the cause of this manipulation.

Participants were assigned the role of moderator, but then also asked to discuss pseudonymously amidst a group of others assigned the same social role. SIDE (Postmes et al., 1998) predicts that individuals in such situations would default to a stronger social identity; similarly, the Proteus Effect is generally assumed to be stronger when participants embody — rather than merely see — an assigned avatar (Yee & Bailenson, 2009). Thus, even if the assignment of a moderator role did cause attitudinal changes, this analysis cannot distinguish between changes caused by the role and changes caused by the following performance of that role.

Rather, the goal of this analysis is best understood as an effort in justifying additional hypotheses about the outcomes of the task. In particular, the notion that the ease of the task may have been driven by the need of moderators to be seen acting lends itself to empirical experimentation. Accordingly, a follow-up experiment was designed to test the specific arguments made here — that telling participants they were being asked to participate “as moderators” altered their behavior by causing them to incorporate a moderator perspective into their decision making, which manifested by driving participants to rate the warning punishment that would specifically show them to be doing something — on a broader scale and provide conclusive evidence of this identity manipulations’ effects.

Chapter 7: A Follow-Up Study

The thematic analysis in Chapter 6 served to identify that some number of participants explicitly or implicitly framed their positions in ways that emphasized a moderator's perspective. However, that analysis could not identify if this was an atypical result or a common one, or establish that it was the prompt's demand for participants to imagine themselves as moderators that drove this effect. Accordingly, a follow-up study was designed to further test these hypotheses. The goals of this study were twofold. The first was to investigate the extent to which prompting participants to act "as moderators" may have influenced their behavior outside of any effect of avatars themselves as observed in the prior analysis. Based on the earlier push for moderators to take action, it was believed that participants specifically acting as moderators may have felt more need to "do something" about even mild misbehavior. Accordingly, it was hypothesized that individuals asked to act as moderators would rank potential punishments differently than those given no such instruction:

H1: Individuals asked to rank punishments for mild misbehavior simply based on their own preferences or, alternatively, while imagining themselves as a community moderator will rank punishments differently.

This hypothesis is noteworthy in that, while moderator was assigned as an identity trait, it is not directly associated with an "avatar" as required by the Proteus Effect. In the initial study, where groups were assigned to act as moderators together, the manipulation more clearly aligns with what SIDE theory considers to be salient group norms.

However, in this follow-up, where individuals lack a group context, this is not the case.

In setting out the Proteus Effect, Yee and Bailenson (2007) argue that group norms and individual identity cues should be seen as distinct characteristics, with the former driving SIDE and the latter driving the Proteus Effect, though both are hypothesized to be influenced by similar factors and share causes. In practice, an individual assigned to a group identity with no local group to exercise norms fits more in line with what Yee and Bailenson discuss as an individual identity cue. If the effects described here happen regardless of the actual presence of a group, it would suggest that the line between individual identity cues and group norms is not as clearly defined as Yee and Bailenson state. Instead, the overarching language of identity cues without artificial boundaries between individual and group traits may help to avoid an unnecessary and illusory distinction.

The second goal of this study was to examine if engaging in discussions, where participants acted out the role assigned to them, may have strengthened identity manipulation effects. This hypothesis was driven from the basic format of Proteus Effect studies, which rely not on assigning individuals an avatar, but on having them use it. Yee and Bailenson (2009) confirm that participants using an avatar reported stronger effects than those merely assigned one. In the study presented in Chapters 3–6, participants were not only given an instruction to adopt a moderator’s identity, but also asked to act out the part socially for a period of time. The manipulations used in this follow-up, however, only involved completing the ranking task without any discussions. While participants can be asked to act out the identity without being asked to do so for longer than a few seconds or to engage with others, the discussants in the previous study clearly did so to a

much larger extent. Likewise, the discussion was a social activity. Van Der Heide et al. (2013) argue that it is the absence of a social environment where participants can influence others through their roles that led to a weaker Proteus Effect. The assumption that a salient identity may become more powerful when shared across an anonymous social group calls back SIDE's (Postmes et al., 1998) emphasis on shared group identities acting as a powerful identity cue in pseudonymous occasions. Accordingly, it was hypothesized that, without the group discussion to facilitate shared identity and role taking, the overall effect of the projected identity in this case would be weaker. As such:

H2: The difference between those asked to rank punishments neutrally will be more pronounced when compared to those who ranked punishments as moderators and completed a group discussion prior to ranking than when compared to those who ranked the punishments without prior discussion.

Methods

This study consisted of a short online experimental survey using materials adapted from the initial discussion experiment. Participants were asked to complete the same ranking task from the previous experiment, though without any accompanying group discussion. Participants were randomly asked either to complete the ranking while pretending they were a community moderator, or simply asked to indicate their preferences.

Participants

155 participants were recruited from the Boston University College of Communication SONA research participant pool in exchange for course credit. This

procedure was identical to that used in the prior study, though no participant who had participated in the first study was allowed to complete the follow-up. A comparison dataset from the previous discussion study was drawn. This group ($n = 113$) consisted of all participants recruited for that study from the SONA research participant pool to maximize comparability. This comparison sample served to test the effects of H2, that the presence of the discussion would lead to stronger identification with a moderator perspective.

Of the new participants, three participants who took less than a minute and four participants who took longer than a day to complete the study were removed from the study due to data quality concerns. It was estimated that it would be quite difficult to read and answer the entire questionnaire within a minute. The upper bound was determined based on the observation that no participants took shortly more or less than a day, so this cutoff allowed for the easy removal of those who would almost certainly have completed the study over multiple sessions. Another 29 participants who did not complete the full procedure were removed from the data set, leaving a final sample of 119 participants recruited for this study, compared to 113 from the prior study. The final newly recruited participants consisted of 66 in the neutral prompt condition and 53 in the moderator prompt condition. Participants had a mean age of 20.4, and participants ranged between 18 and 30 years old; 98 were female, while the other 21 identified as male. Participants primarily identified as Asian ($n = 57$) or white ($n = 50$), with 8 participants identifying as Hispanic, Latino or Spanish Origin, 3 Middle Eastern or North African, 2 Black or African American, 2 Other, and 2 preferring not to say. The demographics of the

comparison discussion group were as reported in Chapter 3. These participants had a mean age of 21.0 years old, with participants between 18 and 27 years of age. In terms of gender, 99 participants were female, 12 were male, 1 was nonbinary and the last indicated they preferred not to answer the question; 62 participants identified as Asian, 43 as White, 8 as Hispanic, Latino or Spanish Origin, 4 preferred not to say, 3 identified as Middle Eastern or North African, and 3 identified as Black or African American.

Materials

Participants were initially asked to complete a short consent form. Following this they were asked to complete a variation on the ranking task used in the discussion study, then fill out the lone psychometric scale (endorsement of aggressive norms) from the prior discussion study. Finally, participants were asked to complete demographic questions consisting of age, gender, race or ethnicity and prior instant messenger experience.

Discussion prompts. Participants were automatically randomly assigned to either a moderator prompt or neutral prompt condition. Both prompts were based on the initial discussion prompt, though did not ask participants to discuss the topic. The prompt consisted of two paragraphs, the first of which was the same for all participants:

Online communities are spaces where anonymous individuals can participate in a variety of discussions on a wide variety of topics, using text, audio and video to communicate over the internet. These communities exist on a number of social media and instant messenger platforms.

Following this, participants in the moderator condition were asked:

I would like you to pretend you are a content moderator for one such community. That is to say, pretend you are in charge of handling misbehavior within the community. A member of the community is routinely mocking or making fun of new users for being unfamiliar with community norms. As moderators, you have five options you can choose to use in response to this misbehavior.

While participants in the neutral condition were instead asked:

Pretend that a member of such a community is routinely mocking or making fun of new users for being unfamiliar with community norms. For these purposes, you should consider five possible consequences for this behavior.

Participants were then asked to rank the 5 response options of increasing severity, used in earlier discussion study (i.e., no punishment, an apology, a warning, a weeklong ban, a permanent ban) from most to least appropriate.

Endorsement of aggressive norms. Participants were given the 3-item endorsement of aggressive norms scale used in discussion one from Peña et al. (2009) that was used in the discussion study ($\alpha = .85$ $M = 1.8$, $SD = 1.0$).

Demographics. Finally, participants were asked to indicate their age, gender, race/ethnicity, and prior experience with instant messenger programs.

Results

A one-way ANOVA identified significant differences in endorsement of aggressive norms between the three conditions (moderator, neutral, moderator with discussion task) evaluated in this study ($F = 3.82$, $p < .05$). Those given a neutral prompt ($M = 1.8$, $SD = 1.02$) and moderator prompt ($M = 1.8$, $SD = 1.0$) both reported less

endorsement of aggressive norms than the pooled participants from the discussion task ($M = 2.2$, $SD = 1.2$). However, Tukey's HSD tests identified no significant pairwise differences between groups.

A series of independent samples median tests was used to evaluate the differences between median ranking of each punishment option between the neutral group, moderator group and the pooled sample from the initial discussion study. The data generated in the first study was ordinal. The necessity of recreating the phenomenon by working as closely as possible with the original dataset necessitated working on the same ranking tasks that were used before. This also helped to preserve comparability between groups but prevented more powerful or sophisticated analytical tools from being used. Additionally, this approach inflates the number of comparisons, as no rankings were independent; however, it has the significant advantage of allowing differences in rankings to be understood not only in whether groups ranked the punishments differently, but also which punishments were ranked differently across groups. Additionally, single items of ranked data are able to be treated as ordinal instead of incorrectly identified as interval or ratio data.

Due to the absence of variance in the ranking of "no punishment," no test-statistics could be computed. As such, in this case, results fail to reject the null hypothesis. Likewise, I fail to reject the null hypothesis when examining the rankings for requesting an apology. The null hypothesis cannot be rejected ($\chi^2(2, N = 229) = 0.24, p = .887$). However, with the remaining warning ($\chi^2(2, N = 232) = 11.00, p < .01$), 1 week ban ($\chi^2(2, N = 229) = 16.76, p < .001$) and permanent ban ($\chi^2(2, N = 232) = 6.61, p < .05$)

options, the null hypothesis is rejected indicating that, in the case of the more severe punishments, differences exist between groups. For comparisons of differences in rankings between groups, see Figures 7.1–7.5 below.

Figure 2: Rankings for "No Punishment" Option

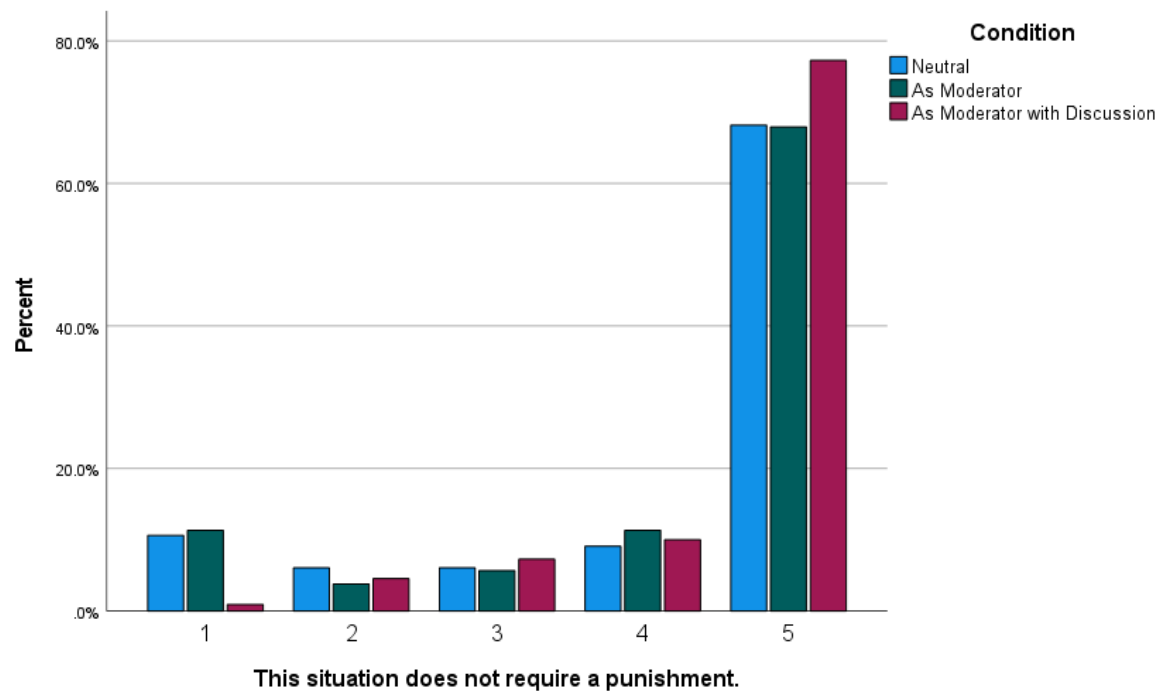


Figure 3: Rankings for "Apology" Option

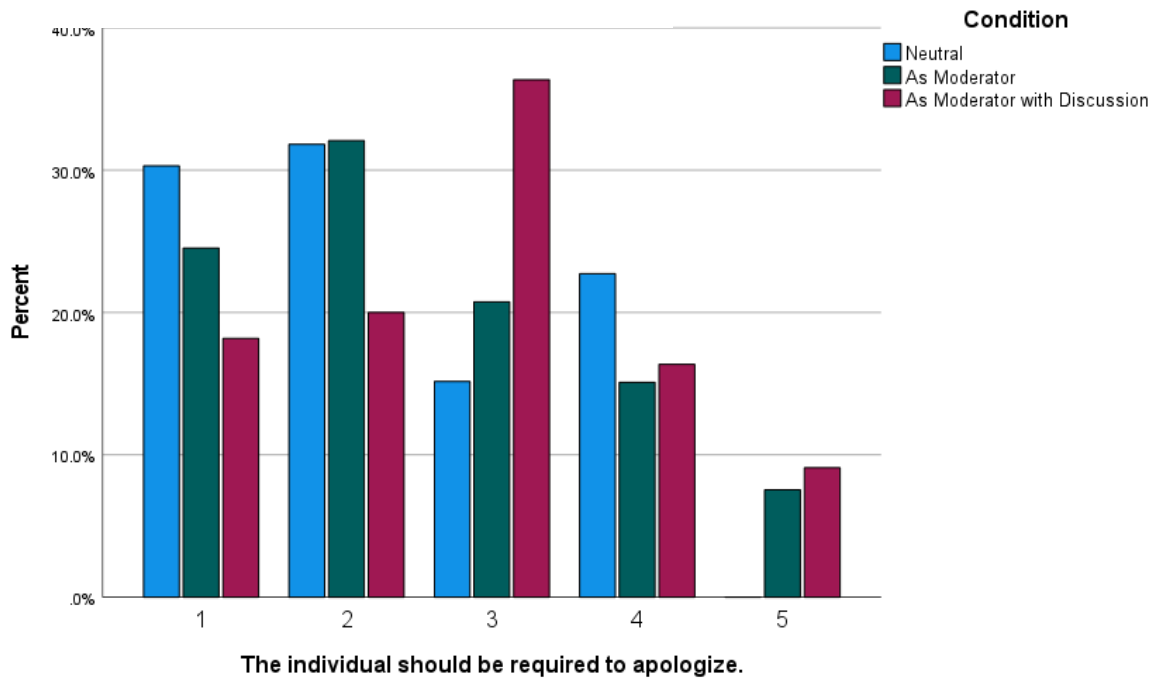


Figure 4: Rankings for "Warning" Option

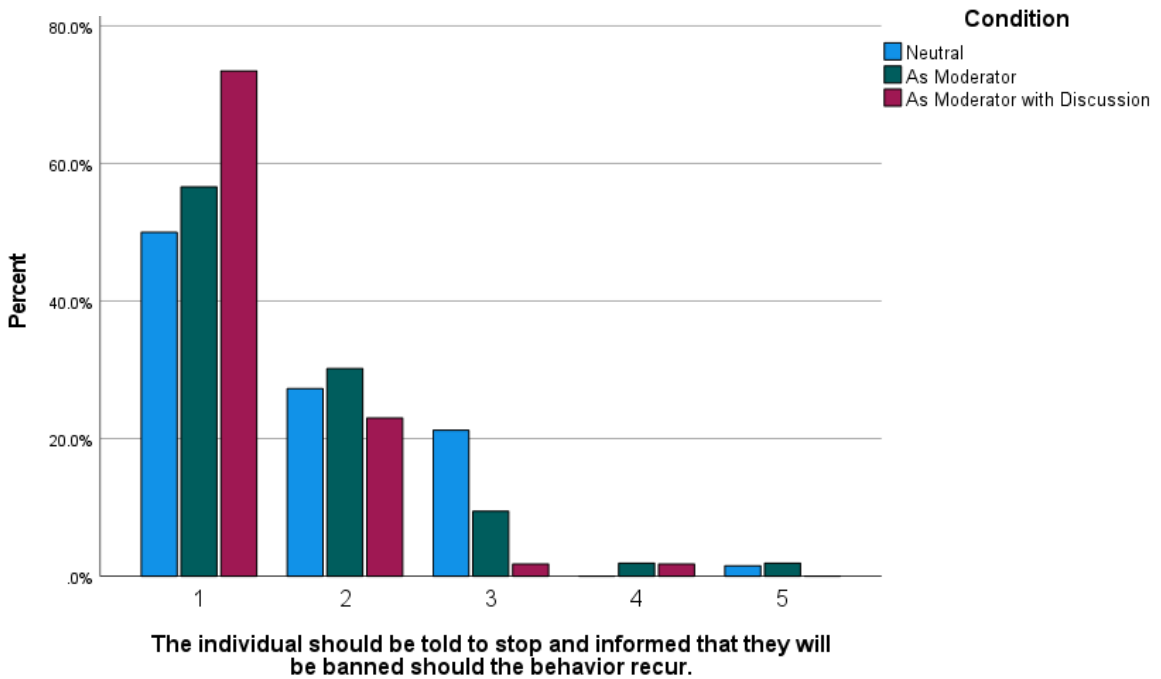


Figure 6: Rankings for "One Week Ban" Option

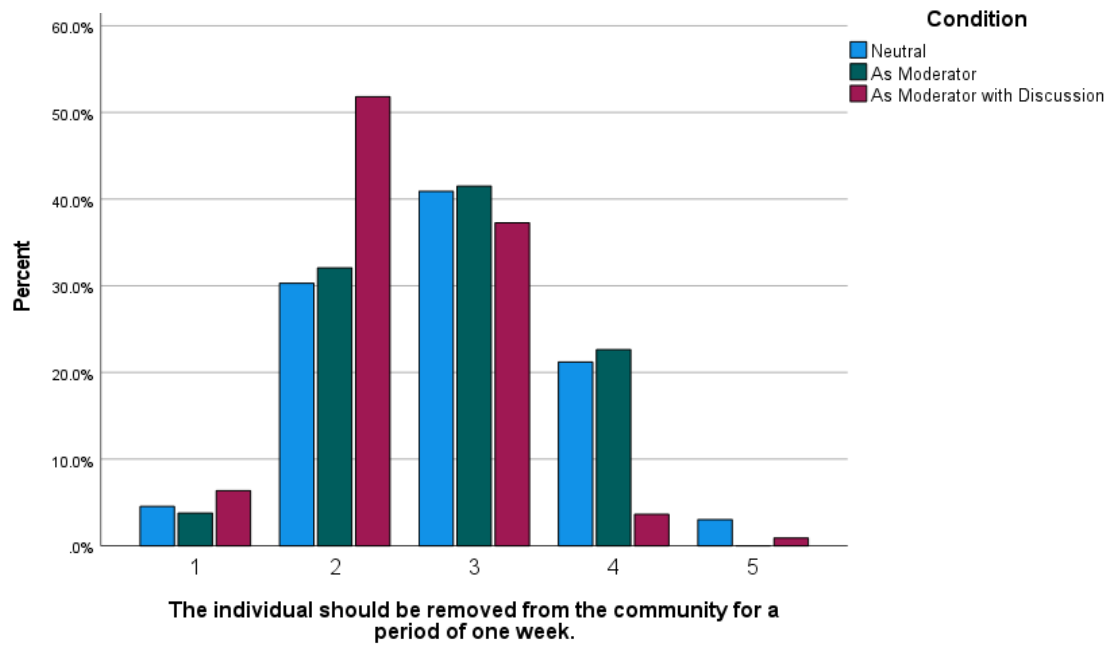
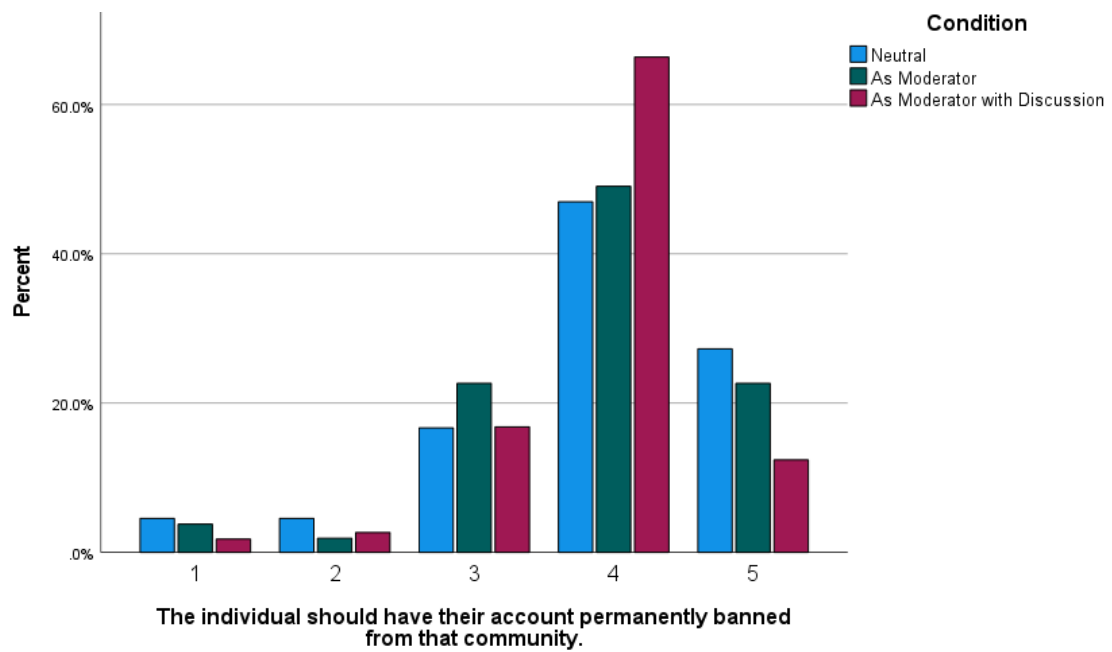


Figure 5: Rankings for "Permanent Ban" Option



In each of these cases, a series of pairwise comparisons was used to identify which groups differed from each other. In each case, Bonferroni corrections were applied due to the multiple simultaneous comparisons. With regards to delivering a warning, significant differences were noted between the neutral and pooled discussion group conditions ($\chi^2(1, N = 179) = 10.05, p < .01$), though not between moderator and neutral conditions ($\chi^2(1, N = 119) = 0.515, p = 1.00$) or moderator and discussion group conditions ($\chi^2(1, N = 166) = 4.71, p = .90$). In particular, the median ranking for those given the neutral prompt was 1.5 vs 1 in the other conditions. In the case of the one-week ban, both neutral ($\chi^2(1, N = 116) = 15.23, p < .001$) and moderator ($\chi^2(1, N = 113) = 7.14, p < .05$) groups differed significantly from the initial discussion group, though not from each other ($\chi^2(1, N = 119) = 0.042, p = 1.00$). This punishment was given a median rank of 3 in both moderation and neutral prompt groups and 2 in the discussion task group. The rankings of the permanent ban option varied only between the neutral and discussion groups ($\chi^2(1, N = 179) = 6.29, p < .05$), but those who were asked to rate as moderators were not significantly different from either the neutral group ($\chi^2(1, N = 119) = 0.334, p = 1.00$) or the discussion group ($\chi^2(1, N = 166) = 2.87, p = 2.71$). The median ranking was 4 for all groups.

Accordingly, while the results did not provide clear evidence of a linear relationship between the three groups, significant differences were largely between the neutral condition and the initial discussion study, with the moderator without discussion condition falling approximately between the two. These results provide support for H2, that the previous study's moderator condition with accompanying discussion would differ

more from the neutral condition than the moderator condition without discussion. Likewise, while the differences between the neutral and moderator prompts were not, themselves, significant, the overall pattern of results included only significant differences between neutral prompts and moderator prompts with accompanying discussion and insignificant differences between moderator with and without discussion groups (except in the case of preference for a one-week ban). This serves as indirect evidence for H1, that differences would exist between those prompted to act as moderators and those not prompted to do so.

Discussion

This second study followed the observation that many participants were observed to have framed their actions as moderators in the first discussion study, in particular selecting uniform answers to the discussion task that emphasized the need of moderators to be seen taking action (and thus accounting for more than 90% of all participants ranking the least severe action option of simply delivering a warning as the first- or second-best answer.) While this was, in a sense, participants simply following instructions, it was not anticipated in advance that simply asking participants to behave as moderators for the sake of framing the discussion topic would engage in a kind of identity play and adopt the social role assigned to them in ways which might alter their behaviors and stated preferences.

To test this empirically, participants answered the question from the first study based solely on their own preferences instead of a discussion, either asked to answer the questions as moderators or neutrally. No significant differences were detected between

groups. However, when comparing both conditions to the initial answers from the discussion study (which was believed to have a stronger identity manipulation as participants were not only asked to think of themselves as moderators but to play the role for an extended discussion) the neutral group differed from the discussion study in rankings of three out of five punishments, while the group prompted to answer as moderators only differed from the discussion study in ranking a single punishment, suggesting that the moderator condition fell roughly between the neutral and discussion conditions, providing indirect support for H1.

Implications

These results lend support to the argument that the prompt to participate “as moderators” effectively served as a second means of imposing an identity on participants, influencing the way they answered questions in an immediately subsequent task. These results, while not stemming from an ‘avatar’ as a particular assemblage of identity characteristics, nonetheless mirror what would be assumed of the Proteus Effect in these circumstances. There is, on one level, a clear technological distinction between assigning participants an avatar that is coded as a moderator and simply asking to imagine themselves to be moderators for a brief task.

On another level, theoretical accounts of these findings seem to align with the hypothesized mechanisms underlying the Proteus Effect, achieved through similar but clearly distinct means. For instance, the most robust explanation of the Proteus Effect (Ratan & Dawson, 2016) suggests the effect operates by tying individuals’ avatar-schemas to self-schemas via avatar embodiment. This is essentially an identical

mechanism to tying individuals' schema for content moderators to their self-schema by having them inhabit the role. This approach favors returning to work by Suler (2001) and Postmes et al. (1997) and treating the Proteus Effect not as a wholly unique phenomenon dependent on particular technological configurations, but one particular means of manifesting certain identity traits while suppressing others, influenced by, but not wholly dependent on the technology. Rather, these results favor integrating the Proteus Effect into the broader anonymity framework outlined in the earlier sections of this dissertation. Under this framework, the results of this study are not necessarily a manifestation of the Proteus Effect as initially construed in Chapter 6. Instead, they can be seen as a distinct manipulation of the same broader underlying phenomena that result in the Proteus Effect in the first place.

Limitations and Future Directions

The chief limitation of this analysis is the limited nature of the data used by necessity to mirror the circumstances in which these phenomena were first observed (a series of items ranked ordinally). This approach, though necessary to compare to the prior study and to maximize the likelihood of recreating the same phenomenon, limited the robustness of analyses that could be conducted and the possible strength of findings. While this analysis provided a meaningful comparison for an ad hoc follow-up study, future research should, accordingly, set out with manipulations in mind that allow for more easily analyzed data to bolster the initial work done here.

One factor that could not be examined clearly in these analyses, but which could be of great interest to future research is that different identity manipulations were

activated at the same time in the initial discussion study and seemed to have achieved independent effects on behavior. This promising finding suggests not only potential avenues of contrasting the strength of differing identity cues directly, but also a wealth of questions as to what happens when identity manipulations are inconsistent with each other. Additionally, by connecting the Proteus Effect to anonymity, it becomes possible to examine what happens in cases where “real life” identity cues are not so hidden as they are in most Proteus Effect studies. For instance, many players of MMO games use voice chat to converse with each other while playing characters that do not resemble the stereotypes associated with their voices. This suggests that offline friends who share group identities may, likewise, enact very different effects from avatars than those whose only salient identity characteristics are those presented by the avatars they use.

Chapter 8: Conclusions

This dissertation began with the notions that anonymity is thought to play a substantial role in the nature of everyday online political talk. However, as Chapter 1 illustrates, anonymity as a concept is often mistreated, left under-defined and oversimplified as a trade-off where all presumed benefits and drawbacks follow from the act of hiding one's identity and becoming deindividuated. Much as SIDE explains offline anonymous behavior better than deindividuation theory by emphasizing what social categories become visible when anonymous (Postmes & Spears, 1998), the broader question of anonymity, Chapter 1 argues, is best addressed by answering what becomes salient when individuals adopt pseudonyms online. This discussion leads to a redefinition of anonymity: *Perceived online anonymity is the totality of identity information individuals perceive to be presented and not presented in computer-mediated contexts, including elements of both perceived self-anonymity — the range and intensity of identity cues an individual presents in online contexts — and perceived other-anonymity — perceptions of others' social presence.*

This definition was used to connect the concept of the Proteus Effect, that individuals adopt behaviors of their avatars in virtual worlds (Yee & Bailenson, 2007) to SIDE's notion of salient group identities influencing behavior (Postmes et al., 1997) and Suler's (2004) conceptualization of dissociative imagination all under the consolidating framework of anonymity. While Yee and Bailenson (2007) do argue that the effect is explicitly different from SIDE due to the emphasis on individual cues rather than group cues, the consolidating definition of anonymity obviates these definitional issues and,

thus, challenges the domain-specificity of the Proteus Effect. In turn, this lends empirical heft to the definition of anonymity presented here as one which reflects the diverse influences anonymity can exert on political discussion. In particular, Peña et al. (2009) found that participants assigned Jedi avatars reported less endorsement of aggressive norms than those assigned “Evil Jedi” avatars.

To justify this connection, it was necessary to show empirically that the Proteus Effect could occur even in text-based environments, and that the factors thought to influence anonymity in the above definition — the salience of identity cues — would serve to highlight the effect’s strength. Accordingly, an experiment was designed using the platform Discord, where participants were asked to imagine themselves as moderators and discuss the rankings of five possible punishments for an individual who routinely mocked new users in a community. In accordance with past Proteus-effect manipulations, participants were randomly assigned either aggressive or unaggressive usernames. These usernames were either accompanied by aggressive or unaggressive profile pictures. Preregistered results were non-significant. That said, a series of alternative analyses found that perceived avatar aggression was related positively to endorsement of aggressive norms and lower group cohesion, particularly in the case where participants were assigned both usernames and visual avatars. That is, the observed effect happened when they were assigned more concurring identity cues. These results show that the Proteus Effect is not platform specific, and that it can translate to the kind of everyday political talk seen on Twitter and Discord and Reddit and a great variety of forums, suggesting that the Proteus Effect can be understood in terms of anonymity as presented

earlier.

A second unanticipated identity manipulation was observed while conducting this study. Rather than only internalizing their account details, it seemed participants adopted the perspective of moderators in the way they argued and, potentially, the conclusions they reached about what were the best rankings of punishments. A qualitative analysis of the discussion transcripts revealed that a minority of participants explicitly (“as a mod...”) or implicitly (e.g., frames of duty or obligations to platform owners) revealed a commitment not just to identifying the best punishments, but doing so as moderators. An online follow-up experiment confirmed that participants in the first study (who were asked to imagine themselves as moderators and discuss appropriate punishments) adopted a different pattern of punishments — ranking the performative option of warning misbehavers more highly — than those merely asked to rank the punishments while acting as moderators or those asked to simply indicate their punishment preferences without any prompted role. Specifically, while the initial study’s participants (asked to imagine themselves as moderators and discuss the issue) varied significantly from both the group asked to describe themselves as moderators and the group given no prompt, the differences between the discussion group were larger when compared to the group given no prompt. These findings suggest that asking participants to imagine themselves as moderators triggered a salient identity cue, with this outcome furthered when then also asked to embody it through group discussion. These results are essentially identical to the Proteus Effect.

However, it is difficult to describe the instruction to act as a moderator as an

avatar effect. This is doubly true in cases where no discussions occurred, and participants were not provided any means of embodiment beyond their own self-concepts. These findings suggest that avatars are not necessary to trigger the kind of identity manipulations that the Proteus Effect causes: simply making traits salient can be enough.

Anonymity Can Consolidate Existing Theory

So, what does it all mean? This research serves to justify empirically a novel definition of anonymity emphasizing, on the one hand, salient identity cues as the primary influence on influencing how individuals perceive themselves, and, on the other hand, social presence as the primary influence on how they perceive others. The present work — an initial example to showcase the utility of this consolidating framework — focused more on self-anonymity than other-anonymity, though measures of social presence were included in studies.

Initial results have been promising. The redefinition of anonymity provides a space for theoretical consolidation of SIDE, disinhibition, and the specific theory of the Proteus Effect using the perceptual language of salient identity cues. This simultaneously serves to challenge the boundaries and arrangements of these theories, and to fit them into a practical framework. Chiefly, SIDE seems to trigger similar effects via salient group identity cues to those observed in the Proteus Effect. Yee and Bailenson (2007) identify this similarity, but take only preliminary steps to resolve it, instead delineating SIDE to purely the influence of groups present at a given moment, while ascribing individual identity cues to be a distinct category. However, SIDE does not conceptualize group identity merely as shared salient traits among those present together, but salient

traits tied to individuals' social identities. Even were a clear distinction to be made between group traits (tied to social circumstances) and individual ones (tied to personal characteristics) it is clear that avatars do not reflect either specifically group or individual traits and these may intersect. Likewise, as the moderator manipulation shows, avatars are not strictly necessary to trigger identity manipulations based on individual traits.

To resolve this dilemma, two possible approaches present themselves. First, one could expand on the initial distinctions posed by Yee and Bailenson to create a more extensive typology of identity manipulations. However, there is little a priori reason to believe such an approach would be meaningful. Instead, this work suggests that a wide variety of these effects can be consolidated under the label of identity cues and anonymity. Theoretically, this approach suggests that findings drawn from different contexts and theories should be used to inform each other until theoretical distinctions can be seen empirically. For example, the role of identification has been noted in studies of the Proteus Effect; is there any reason to presuppose these should not extend to group identity manipulations or individual identity manipulations that are not tied to avatars? Is there any reason that identification with assigned identities may not equally influence SIDE?

This is not to diminish the importance of avatars — complex assemblages with a wide variety of traits (Banks, 2018) — or the role of shared group identities or archetypal images. Rather, each of these distinct phenomena are interesting and important cases, and the technical opportunities they offer to influence what identity cues individuals perceive themselves giving off are important and worthy of study. By testing the specific role of

particular contexts which can trigger similar behavioral changes future research can better discriminate and identify theoretically meaningful differences that do not translate over these contexts. This would allow for clearer identification of the unique role of factors such as avatars.

By fitting each of these within the consolidating framework of anonymity, it becomes possible to examine these theories in context of each other and to prod the boundaries of each. This, in turn, creates room for further advancing these individual theories. For instance, the studies here suggested that multiple distinct types of identity cues — names, pictures, text prompts containing the assignment of a role — exerted distinct influences on participants. Under the umbrella of self-anonymity, future work can contrast different types of identity cues to better understand which exert greater influence and how these influences interact. Preliminary evidence suggests that social presence — perceptions of others' anonymity — may have exerted a much stronger influence on group cohesion than self-anonymity did. SIDE posits that group identity matters as an identity cue, and social presence will undoubtedly influence the salience of these individual identity cues.

Anonymity Need Not Be a Trade-Off

The redefinition of anonymity investigated here provides a clear research agenda for investigating the role anonymity plays in online political discussion. These findings suggest that anonymity can have diverse effects, rather than simply presenting all effects on all occasions. Anonymity, when it leads to a presentation that is fundamentally aggressive, can lead individuals to behaving more aggressively. This, in turn, can be

anticipated to lead to outcomes such as incivility and harassment, but also to more willingness to call out immoral conduct. Other arrangements of anonymity could easily emphasize shared communal identities and values as Walther et al. (2015) suggest occurs when communication becomes hyperpersonal. Following past Proteus Effect research, it seems likely that different arrangements of salient identity cues could lead individuals to adopting gendered behavior, to become more or less assertive or to any number of other effects.

Traditionally, the trade-off perspective of anonymity has argued that the benefits of anonymity entail accepting its drawbacks. This redefinition, and the empirical results observed here, suggest that this need not be the case. Not all anonymity is created equally and the particular configuration of anonymity and how individuals use it can be modeled more effectively in research to build clearer and better causal models.

It is tempting to translate these recommendations to policy. However, there are a number of risks in doing so. As Russell (2020) argues, the range of anonymity as a tool of self-expression can allow individuals with marginalized identities to enact race, gender, etc. online in important ways helpful to identity formation. Restricting anonymous identity performance to top-down instantiations of prosocial effects could easily hamper these activities or alienate platform users. Likewise, it is not necessarily easy to translate specific technical manipulations into user perceptions of those manipulations. Perceptions of avatars can likely be influenced by beliefs about how those avatars are constructed, which can, in turn, be influenced by policies and affordances platforms provide. Additionally, it can be difficult to translate specific identity cues into

policy recommendations. Aggression need not be a bad thing in all instances, nor does a lack of aggression automatically constitute a net good. Other manipulations may exert similar complicated influences. More narrowly scoped instances of political talk which aim to evoke a particular atmosphere or conversation dynamic might benefit from these manipulations in ways that large platforms do not. At the same time, even large platforms should be aware that anonymity should not be anticipated to lead to deleterious effects in all cases.

Anonymity is Many and Varied

On November 10th, 2022, someone took advantage of Twitter's newly revised verification system to create an anonymous account. The account presented itself as the official Twitter presence of pharmaceutical giant Eli Lilly. They tweeted that: "We are excited to announce insulin is free now" (Shimunov, 2022). Looking realistic, this statement garnered thousands of responses in the six hours it took Twitter to remove the tweet. Eli Lilly's stock price plummeted, and the company, as well as other producers of insulin, lost billions of dollars in market cap in the space of an afternoon as this tweet sparked discussion on the high prices charged for insulin, a cheaply produced life-saving medication (Adams, 2022).

Eli Lilly's shareholders would potentially classify this act as a kind of toxic disinhibition; an individual using anonymity to engage in wildly destructive behavior thanks to the obscuring and deindividuating effects of hiding one's name. From another perspective, this was instead a distinctly sophisticated political performance of anonymity. The tweeter not only hid their real name, but actively created another identity,

which granted the message its power to challenge corporate practices. In this sense, anonymity in the Eli Lilly incident was hardly deindividuating. The individual's anonymity came from presenting a series of highly visible identity cues: Eli Lilly's name, its corporate logo, and a small white checkmark on a blue background. Together, these things enabled the anonymous satirist to enact the role of Eli Lilly: not to take on the traits of the company, but to critique them.

Why mention this example? The incident, one of many similar hoaxes which occurred in the same short time period, illustrates the complex roles anonymity can play in causing political talk online, and showcases that anonymity itself can drive significant and very real consequences. While the language of salient identity cues is apt to describe the events, the specific anonymous performance here was bent toward a use fundamentally distinct from the ways in which anonymity has been used in this dissertation. The issue of the internet and informal everyday political talk is much larger than the narrow question of anonymity. Even then, the role of anonymity is broader and more convoluted than could adequately fit within a single series of studies.

Future research should take this dissertation's revised definition of anonymity as a launching point to further pursue the many and varied anonymities that exist together online.

Appendix A: Discussion Group Script

The following is the full script for the discussion groups that took place during the study described in Chapters 3–5 is below. Text in parentheses and italicized is instructions the chat moderator was to follow where exact wording could not be provided, while all other text was copied and pasted into the chat during each discussion group. Additionally, chat moderators answered questions if asked directly. All text is exactly as typed in though session numbers and participant usernames would vary by session, and the email used and survey link have been omitted. No formatting has been altered.

(The researcher will greet individual participants as they sign in. These times may be somewhat spread out due to the technical elements of this design. Once everyone has logged in, the researcher will begin.)

Hello everyone! Everyone seems to have logged in successfully. Could everyone type something into the textbox near the bottom of your window now to let me know that the program is working?

(The researcher will wait for everyone to type something into the chat window to confirm that the software is working successfully. If there are issues, the researcher will attempt to troubleshoot.)

Great, it looks like everything is working! Does anyone have any questions before we begin?

(The researcher will briefly wait for any comments or questions.)

Okay. The program we're using right now is called Discord. Discord is widely used by online communities of all sizes for both text and video chat. For the discussion today, I

would like you to pretend you are a content moderator on one such community. That is to say, pretend you are in charge of handling misbehavior within the community. A member of a community is routinely mocking or making fun of new users for being unfamiliar with community norms. As moderators, you have five options you can choose to use in response to this misbehavior.

These options are:

1. This situation does not require a punishment.
2. The individual should be required to apologize.
3. The individual should be told to stop and informed that they will be banned should the behavior recur.
4. The individual should be removed from the community for a period of one week.
5. The individual should have their account permanently banned from that community.

I would like you to first discuss the merits and drawbacks of each punishment in this situation, before ranking them from most to least appropriate for this situation. This conversation should take approximately 15 to 20 minutes and will be stopped after 20 minutes. There are no right or wrong answers to the scenario

Does anyone have any questions, or are you ready to begin?

(The researcher will briefly wait for any comments or questions.)

Okay, please feel free to begin!

(The researcher will wait and observe as the participants discuss, answering questions or resolving technical difficulties as needed. After the group has decided, or twenty minutes have passed, the researcher will respond.)

Okay, good job everyone, unfortunately we're out of time. Could you please now summarize your rankings for each punishment?

(Or, if the participants reach a conclusion early or simply stop talking in chat)

Okay, it looks like you've reached a conclusion. Does anyone else have anything they'd like to add?

(been a bit quiet, instead of reached a conclusion if they're quiet.)

(The researcher will allow the group to post or explain that they haven't reached a consensus.)

Great, thank you!

That concludes the discussion part of this study. I will now post a link to an external survey. The survey will ask you for the username you used during the discussion and your session number. Your session number is ****#####****. Please remember both.

Additionally, when you have finished, please do not discuss this study with anyone, as data is still being collected. If you have questions or concerns, you can reach out to EMAIL ADDRESS or any of the contact methods mentioned on the consent form you signed.

Here is the link to the survey: (Qualtrics Survey Link)

Your usernames are the names you used in the chat, that is _____, _____ or _____

Appendix B: Study 1 Questionnaire

The full study 1 questionnaire is below. Question order was randomized as described in Chapter 3. Descriptions of scale, as well as sources, are included in parentheses.

(Instructions)

Thank you very much for participating in this study! In the following questionnaire, you will be asked to enter several questions concerning your experiences during the discussion. This questionnaire should take approximately 5 minutes to complete, after which you will be finished with the study.

Please enter the username of the account you used during the discussion here:
At the end of the discussion you were instructed to copy and paste a numerical code.
Please enter the room code here:

(Short et al., 1976; Perceived Salience of Interactant) For each of the pairs of words below, please circle the number that best describes your evaluation of the discussion experience.

Dead	1	2	3	4	5	6	7	Lively
Unsociable	1	2	3	4	5	6	7	Sociable
Impersonal	1	2	3	4	5	6	7	Personable
Insensitive	1	2	3	4	5	6	7	Sensitive
Remove	1	2	3	4	5	6	7	Immediate
Unemotional	1	2	3	4	5	6	7	Emotional
Unresponsive	1	2	3	4	5	6	7	Responsive

(Lowden & Hostetter; 2012, association and mutual awareness) Please indicate the extent to which you agree with each of the following statements about the discussion. (Strongly Disagree to Strongly Agree)

- I felt comfortable conversing through this medium in the meeting.
- I felt comfortable interacting with other participants in the meeting
- I felt comfortable participating in the meeting discussion
- I felt that my point of view was acknowledged by other participants in the meeting
- I was able to form distinct individual impressions of some meeting participants

(Downs et al., 2019; identification) Please indicate the extent to which you agree with each of the following statements concerning the account you used to represent you during the discussion. (Strongly Disagree to Strongly Agree)

(Liking) I like this account.

(Liking) I dislike this account.

(Liking) I have positive feelings toward this account.

(Liking) I feel like this avatar is interesting.

(Embodiment) I felt like I was inside this account during the discussion.

(Embodiment) During the discussion I was transported into this account.

(Embodiment) During the discussion, it felt as if I was this account.

(Embodiment) During the discussion it was as if I'd become one with the account.

(Embodiment) When discussing, it felt as if the account's body became my own.

(Embodiment) During the discussion it was as if I acted directly through this account.

(Peña et al., 2009; group cohesion and endorsement of aggressive norms)

(Group Cohesion) To what degree did you feel that you are really a part of your group?

(Not at all < > Extremely)

(Group Cohesion) How does your group compare with other groups on the way people get along together? (Much worse < > Much better)

(Group Cohesion) How does your group compare with other groups on the way people helped each other on the task? (Much worse < > Much better)

(Endorsement of Aggressive Norms) Please indicate the extent to which you agree with each of the following statements concerning the account you used during the discussion.

I would make fun of someone if I ever participated in online discussions using this account.

Using this account, for me to make fun of someone would be good

Others using accounts like this would often mock others.

(Adapted from Nowak & Ruah, 2006; manipulation checks) Please indicate the extent to which you agree with each of the following statements about the avatar and/or username you used during the discussion. (Strongly Disagree < > Strongly Agree)

This account seemed aggressive

This account seemed feminine

This account seemed masculine

This account seemed intelligent

This account seemed reliable

This seemed like an account I would see someone use on an instant messenger

(Peña et al., 2009; Open Ended awareness check)

What did you think the experiment was about?

What did you think the researchers' hypothesis was (i.e., what did you think they were looking for, trying to study, etc.)?

If you were suspicious about the goals of the experiment, when did you figure them out?

(Demographics)

Please enter your age below (in years).

Please indicate your gender.

Male

Female

Other (please indicate preferred identification)

Prefer not to say

Which categories describe you? Please select all that apply to you.

American Indian or Alaska Native

Black or African American

Asian

Native Hawaiian or Other Pacific Islander

White

Hispanic, Latino or Spanish Origin

Middle Eastern or North African

Other

Prefer not to say

Which of the following instant messaging programs have you used in the past? (Please check all that apply.)

Discord

Slack

WhatsApp

WeChat

Line

IRC

AOL Instant Messenger

Signal

Other

None of the Above

Appendix C: Alternate Multilevel Model Specifications

The following tables depict alternate specifications for multilevel models described in Chapter 4, specifically, these are identical to the preregistered analyses reported in that chapter, except that the experimental manipulation has been replaced by the manipulation check question. Dependent variables are as specified in table labels.

Table 11: Alternate Multilevel Models for Agreement

	Empty model	Independent variables only	With interaction term	With psychometrics	With identification interaction	With demographic variables
Variables		Estimate	Estimate	Estimate	Estimate	Estimate
Perceived Account Aggressiveness		-0.02	-0.04**	-0.03*	-0.03*	-0.03*
Avatar		-0.06	-0.06	-0.06	-0.06	-0.07*
Perceived Account Aggressiveness*Avatar			0.04*	0.04	0.04	0.04
Salience				-0.04*	-0.04	-0.04
Association				0.02	0.02	0.02
Liking				0.01	0.01	0.01
Embodiment				0.00	0.00	0.00
Liking*Aggression					0.00	0.00
Embodiment*Aggression					0.00	0.00
Age						0.00
Gender (male is reference group)						-0.02
Race (all nonwhite is reference group)						0.01
Recruitment Method						-0.05
Summary Statistics						
τ_{00}	0.00	0.00	0.00	0.00	0.00	0.00
σ^2	0.03	0.04	0.03	0.03	0.04	0.03
ICC	0.07	0.03	0.03	0.00	0.00	0.01
<i>Level-1 pseudo R²</i>	NA	-0.02	0.07	0.01	-0.01	0.04
AIC	-52.39	-45.31	-44.10	-24.14	-9.34	8.55

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

Table 12: Alternate Multilevel Models for Disagreement

	Empty model	Independent variables only	With interaction term	With psychometrics	With identification interaction	With demographic variables
Variables		Estimate	Estimate	Estimate	Estimate	Estimate
Perceived Account Aggressiveness		0.01	0.01	0.01	0.01	0.01
Avatar		0.03	0.03	0.04	0.03	0.04
Perceived Account Aggressiveness* Avatar			-0.01	-0.01	-0.01	-0.02
Salience				0.00	0.00	0.00
Association				-0.034*	-0.032*	-0.04**
Liking				0.00	-0.01	-0.01
Embodiment				0.01	0.01	0.01
Liking*Aggression					0.01	0.01
Embodiment* Aggression					-0.01	-0.01
Age						0.00
Gender (male is reference group)						-0.07
Race (all nonwhite is reference group)						-0.03
Recruitment Method						-0.08
Summary Statistics						
τ_{00}	0.01	0.01	0.01	0.01	0.01	0.01
σ^2	0.01	0.01	0.01	0.01	0.01	0.01
ICC	0.44	0.42	0.42	0.48	0.53	0.43
<i>Level-1 pseudo R2</i>	NA	-0.01	-0.02	0.08	0.14	0.04
AIC	-133.70	-123.00	-116.53	-94.66	-82.09	-53.13

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

Table 13: Alternate Multilevel Models for Elaboration

	Empty model	Independent variables only	With interaction term	With psychometrics	With identification interaction	With demographic variables
Variables		Estimate	Estimate	Estimate	Estimate	Estimate
Perceived Account Aggressiveness		-0.01	-0.01	-0.01	-0.01	0.01
Avatar		0.00	0.00	-0.01	-0.01	0.01
Perceived Account Aggressiveness* Avatar			-0.01	0.00	0.00	-0.02
Salience				0.02	0.02	0.03
Association				-0.02	-0.02	-0.05
Liking				0.00	0.00	0.01
Embodiment				-0.01	-0.01	-0.01
Liking*Aggression					0.00	0.00
Embodiment*Aggression					0.00	0.00
Age						0.00
Gender (male is reference group)						-0.14*
Race (all nonwhite is reference group)						-0.07
Recruitment Method						0.07
Summary Statistics						
τ_{00}	0.02	0.02	0.02	0.01	0.01	0.01
σ^2	0.03	0.03	0.03	0.04	0.04	0.04
ICC	0.31	0.33	0.33	0.25	0.25	0.11
<i>Level-1 pseudo R²</i>	NA	0.01	0.00	-0.08	-0.10	-0.18
AIC	-21.76	-11.12	-5.52	15.92	30.12	42.61

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

Table 14: Alternate Multilevel Models for Endorsement of Aggressive Norms

	Empty model	Independent variables only	With interaction term	With psychometrics	With identification interaction	With demographic variables
Variables		Estimate	Estimate	Estimate	Estimate	Estimate
Perceived Account Aggressiveness		.17*	0.31***	0.26**	0.26**	0.26**
Avatar		0.18	0.18	0.15	0.18	0.25
Perceived Account Aggressiveness* Avatar			-0.32*	-0.31*	-0.29*	-0.3e*
Salience				0.08	0.07	0.08
Association				-0.31*	-0.32*	-0.33**
Liking				-0.12	-0.05	-0.07
Embodiment				0.16	0.15	0.16
Liking*Aggression					-0.06	-0.04
Embodiment* Aggression					0.09	0.08
Age						0.01
Gender (male is reference group)						-0.54†
Race (all nonwhite is reference group)						0.06
Recruitment Method						0.25
Summary Statistics						
τ_{00}	0.09	0.06	0.00	0.00	0.00	0.00
σ^2	1.36	1.33	1.33	1.27	1.26	1.31
ICC	0.06	0.04	0.00	0.00	0.00	0.00
<i>Level-1 pseudo R2</i>	NA	0.02	0.02	0.06	0.07	0.03
AIC	441.15	438.76	434.79	426.93	431.69	421.79

† $p < .1$ * $p < .05$, ** $p < .01$, *** $p < .001$

Table 15: Alternate Multilevel Models for Group Cohesion

	Empty model	Independent variables only	With interaction term	With psychometrics	With identification interaction	With demographic variables
Variables		Estimate	Estimate	Estimate	Estimate	Estimate
Perceived Account Aggressiveness		-0.17**	-0.06	-0.01	-0.02	-0.01
Avatar		-0.08	-0.08	-0.06	-0.03	-0.01
Perceived Account Aggressiveness* Avatar			-0.24	-0.18	-0.17	-0.17
Salience				0.380***	0.39***	0.39***
Association				0.389***	0.36***	0.35***
Liking				0.12	0.14	0.16988†
Embodiment				0.02	-0.03	-0.03
Liking*Aggression					-0.114419*	-0.13*
Embodiment* Aggression					0.01	0.01
Age						-0.01
Gender (male is reference group)						-0.05
Race (all nonwhite is reference group)						-0.09
Recruitment Method						-0.40
Summary Statistics						
τ_{00}	0.21	0.24	0.21	0.05	0.05	0.04
σ^2	1.15	1.07	1.07	0.70	0.68	0.72
ICC	0.15	0.19	0.16	0.07	0.07	0.06
<i>Level-1 pseudo R2</i>	NA	0.07	0.07	0.39	0.41	0.37
AIC	430.70	428.11	426.71	361.35	364.53	361.20

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

Bibliography

- Adams, B. (2022, November 15). *Eli Lilly pulls Twitter ads after Blue Check Fallout: Report*. Fierce Pharma. Retrieved November 18, 2022, from <https://www.fiercepharma.com/marketing/eli-lilly-pulls-twitter-ads-after-blue-check-fallout-report>
- Ammari, T., Schoenebeck, S., & Romero, D. (2019). Self-declared throwaway accounts on Reddit: How platform affordances and shared norms enable parenting disclosure and support. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 1–30.
- Anonymous. (1998). To reveal or not to reveal: A theoretical model of anonymous communication. *Communication Theory*, 8(4), 381–407.
- Asenbaum, H. (2018). Anonymity and democracy: Absence as presence in the public sphere. *American Political Science Review*, 112(3), 459–472.
- Banakou, D., Groten, R., & Slater, M. (2013). Illusory ownership of a virtual child body causes overestimation of object sizes and implicit attitude changes. *Proceedings of the National Academy of Sciences of the United States of America*, 110(31), 12846–12851.
- Banks, J. (2018). *Avatar, Assembled*. New York: Peter Lang.
- Bem, D. J. (1972). Self-perception theory. In *Advances in experimental social psychology* (Vol. 6, pp. 1–62). Academic Press.
- Bernstein, M., Monroy-Hernández, A., Harry, D., André, P., Panovich, K., & Vargas, G. (2011). 4chan and/b: An Analysis of Anonymity and Ephemerality in a Large Online Community. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 5, No. 1).
- Beyea, D. (2019). *Avatar-Based Self-Influence in a Text-Based CMC Environment*. Michigan State University.
- Beyea, D., Ratan, R., Lei, Y., Liu, H., Hales, G., & Lim, C. (2022, May 26–30). *Toward a Clear Definition and Understanding of the Proteus Effect: Examining Modality and Avatar Uncanniness as Moderators* [Paper Presentation]. 72nd Annual International Communication Association Conference.

- Chadwick, A. (2008). Web 2.0: New challenges for the study of e-democracy in an era of informational exuberance. *I/S: A Journal of Law and Policy for the Information Society*, 5, 9.
- Chaffee, S. H. (1991). *Communication Concepts 1: Explication*. Newbury Park, CA: Sage.
- Chester, A., & Gwynne, G. (1998). Online teaching: Encouraging collaboration through anonymity. *Journal of Computer-Mediated Communication*, 4(2), JCMC424.
- Cho, D., & Acquisti, A. (2013). The more social cues, the less trolling? An empirical study of online commenting behavior. In *Twelfth Annual Workshop on the Economics of Information Security (WEIS 2013)*. Carnegie Mellon University. <https://doi.org/10.1184/R1/6472058.v1>
- Conover, P. J., & Searing, D. D. (2005). Studying ‘everyday political talk’ in the deliberative system. *Acta Politica*, 40(3), 269–283.
- Cummings, J. J., & Wertz, E. E. (in press). Capturing social presence: Concept explication through an empirical analysis of social presence measures. *Journal of Computer-Mediated Communication*
- Cummings, J. J., & Wertz, B. (2018). Technological predictors of social presence: A foundation for a meta-analytic review and empirical concept explication. In *Proceedings of the 10th Annual International Workshop on Presence (Prague)*. <http://matthewlombard.com/ISPR/Proceedings/2018/P2018-Cummings%20&%20Wertz.pdf>
- Davies, T., & Chandler, R. (2012). Online deliberation design. In Tina Nabatchi, John Gastil, G. Michael Weiksner & Matt Lehniger (eds.), *Democracy in Motion: Evaluating the Practice and Impact of Deliberative Civic Engagement*. (pp. 103–131). New York: Oxford University Press.
- Diehl, M. (1990). The minimal group paradigm: Theoretical explanations and empirical findings. *European Review of Social Psychology*, 1(1), 263–292.
- Fox, J., Bailenson, J. N., & Tricase, L. (2013). The embodiment of sexualized virtual selves: The Proteus Effect and experiences of self-objectification via avatars. *Computers in Human Behavior*, 29(3), 930–938.
- Friess, D., & Eilders, C. (2015). A systematic review of online deliberation research. *Policy & Internet*, 7(3), 319–339.

- Goffman, E. (1978). *The presentation of self in everyday life*. London: Harmondsworth.
- Greaves, A. (2022). *Welcome to Dorley Hall: The Sisters of Dorley Book One*. [No Publisher]
- Griffith, E., Weiss, K., & Browning, K. (2021, March 23). Discord and Microsoft Said to Discuss Deal That Could Top \$10 Billion—The New York Times. *The New York Times*. <https://www.nytimes.com/2021/03/23/technology/microsoft-discord-deal.html>
- Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
- Haimson, O. L., Dame-Griff, A., Capello, E., & Richter, Z. (2021). Tumblr was a trans technology: the meaning, importance, history, and future of trans technologies. *Feminist Media Studies*, 21(3), 345–361.
- Hayes, A. F. (2006). A primer on multilevel modeling. *Human Communication Research*, 32(4), 385–410.
- Hofmann, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. *Journal of Management*, 24(5), 623–641.
- Hollenbaugh, E. E., & Everett, M. K. (2013). The effects of anonymity on self-disclosure in blogs: An application of the online disinhibition effect. *Journal of Computer-Mediated Communication*, 18(3), 283–302.
- Howard, P. N., Duffy, A., Freelon, D., Hussain, M. M., Mari, W., & Maziad, M. (2011). Opening closed regimes: what was the role of social media during the Arab Spring? Available at SSRN 2595096.
- Jaidka, K., Zhou, A., Lelkes, Y., Egelhofer, J., & Lecheler, S. (2022). Beyond anonymity: Network affordances, under deindividuation, improve social media discussion quality. *Journal of Computer-Mediated Communication*, 27(1). <https://doi.org/10.1093/jcmc/zmab019>
- Jardine, E. (2015). The Dark Web dilemma: Tor, anonymity and online policing. *Global Commission on Internet Governance Paper Series*, (21). https://www.cigionline.org/documents/963/no.21_1.pdf

- Kao, D., Ratan, R., Mousas, C., Joshi, A., & Melcer, E. F. (2022). Audio Matters Too: How Audial Avatar Customization Enhances Visual Avatar Customization. *arXiv preprint arXiv:2202.05315*.
- Kabay, M. E. (1998, March). Anonymity and pseudonymity in cyberspace: deindividuation, incivility and lawlessness versus freedom and privacy. In *Annual Conference of the European Institute for Computer Anti-virus Research (EICAR), Munich, Germany* (pp. 16–8). <https://www.mekabay.com/overviews/anonpseudo.pdf>
- Karpowitz, C. F., & Mendelberg, T. (2014). *The silent sex: Gender, deliberation, and institutions*. Princeton University Press.
- Leavitt, A. (2015). "This is a Throwaway Account" Temporary Technical Identities and Perceptions of Anonymity in a Massive Online Community. In *CSCW'15: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 317–327). <https://doi.org/10.1145/2675133.2675175>
- Lee, J. E. R., Nass, C. I., & Bailenson, J. N. (2014). Does the mask govern the mind?: Effects of arbitrary gender representation on quantitative task performance in avatar-represented virtual groups. *Cyberpsychology, Behavior, and Social Networking*, 17(4), 248–254.
- Leshed, G. 2009. "Silencing the Clatter: Removing Anonymity From a Corporate Online Community." In T. Davies and S.P. Gangadhara (eds.) *Online Deliberation: Design, Research, and Practice*. (pp. 243–251). Stanford, CA: CSLI Publications.
- Lunden, I. (2020). Update: Discord confirms raising \$100M at a valuation of \$7B. *TechCrunch*. <https://social.techcrunch.com/2020/12/17/filing-discord-is-raising-up-to-140m-at-a-valuation-of-up-to-7b/>
- Manikonda, L., Beigi, G., Kambhampati, S., & Liu, H. (2018, July). #metoo through the lens of social media. In *International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation* (pp. 104–110). Springer, Cham.
- Mansbridge, J. (1999). Everyday talk in the deliberative system. In S. Macedo (Ed.), *Deliberative politics: Essays on democracy and disagreement*. (pp. 211–240). Oxford University Press.

- Marwick, A. E., & boyd, D. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, 13(1), 114–133.
- Marx, G. T. (1999). What's in a Name? Some Reflections on the Sociology of Anonymity. *The Information Society*, 15(2), 99–112.
- Massanari, A. (2017). # Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, 19(3), 329–346.
- Matias, J. N. (2019). The civic labor of volunteer moderators online. *Social Media + Society*, 5(2), 2056305119836778.
- Mendelberg, T., Karpowitz, C. F., & Oliphant, J. B. (2014). Gender inequality in deliberation: Unpacking the black box of interaction. *Perspectives on Politics*, 12(1), 18–44. <https://doi.org/10.1017/S1537592713003691>
- Mozur, P. (2018, October 15). A genocide incited on Facebook, with posts from Myanmar's military. *The New York Times*. Retrieved November 20, 2022, from <https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html>
- Moore, A. (2018). Anonymity, pseudonymity, and deliberation: Why not everything should be connected. *Journal of Political Philosophy*, 26(2), 169–192.
- Moore, A., Fredheim, R., Wyss, D., & Beste, S. (2020). Deliberation and identity rules: The effect of anonymity, pseudonyms and real-name requirements on the cognitive Complexity of online news comments. *Political Studies*, 69(1), 45–65. <https://doi.org/10.1177/0032321719891385>.
- Noam, E. M. (2005). Why the Internet is bad for democracy. *Communications of the ACM*, 48(10), 57–58.
- Nowak, K. L., & Rauh, C. (2008). Choose your “buddy icon” carefully: The influence of avatar androgyny, anthropomorphism and credibility in online interactions. *Computers in Human Behavior*, 24(4), 1473–1493.
- Nowak, K. L., (2018). Race & Otherness: The utopian promise and divided reality. In Banks, J. (Es.), *Avatar, assembled: The social and technical anatomy of digital bodies* (pp. 33–42). New York: Peter Lang.

- Clark, O. J. (2020, October 28). How To kill A greek god – A review, critique, and meta-analysis of 14 years of Proteus Effect research. *PsyArXiv*
<https://doi.org/10.31234/osf.io/9rbcs>
- O'Keefe, D. J. (2003). Message properties, mediating states, and manipulation checks: Claims, evidence, and data analysis in experimental persuasive message effects research. *Communication Theory*, 13(3), 251–274.
- Paccagnella, O. (2006). Centering or not centering in multilevel models? The role of the group mean and the assessment of group effects. *Evaluation Review*, 30(1), 66–85.
- Paulhus, D. (2001). Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 61–84). Routledge.
- Pavalanathan, U., & De Choudhury, M. (2015, May). Identity management and mental health discourse in social media. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 315–321). ACM.
<https://doi.org/10.1145/2740908.2743049>
- Peña, J., Hancock, J. T., & Merola, N. A. (2009). The priming effects of avatars in virtual settings. *Communication Research*, 36(6), 838–856.
- 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. (1998). Deindividuation and antinormative behavior: A meta-analysis. *Psychological Bulletin*, 123(3), 238.
- Rainie, H., Anderson, J. Q., & Albright, J. (2017). *The future of free speech, trolls, anonymity and fake news online*. Washington, DC: Pew Research Center.
- Randel, A. E. (2002). Identity salience: A moderator of the relationship between group gender composition and work group conflict. *Journal of Organizational Behavior*, 23(6), 749–766.
- Ratan, R., Beyea, D., Li, B. J., & Graciano, L. (2019). Avatar characteristics induce users' behavioral conformity with small-to-medium effect sizes: A meta-analysis of the proteus effect. *Media Psychology*, 23(5), 651–675.

- Ratan, R., & Dawson, M. (2016). When Mii is me: A psychophysiological examination of avatar self-relevance. *Communication Research*, 43(8), 1065–1093.
- Ratan, R., Rikard, R. V., Wanek, C., McKinley, M., Johnson, L., & Sah, Y. J. (2016, January). Introducing Avatarification: An experimental examination of how avatars influence student motivation. In *2016 49th Hawaii International Conference on System Sciences (HICSS)* (pp. 51–59). IEEE.
- Ratan, R., & Sah, Y. J. (2015). Leveling up on stereotype threat: The role of avatar customization and avatar embodiment. *Computers in Human Behavior*, 50, 367–374.
- Reicher, S. D., Spears, R., & Postmes, T. (1995). A social identity model of deindividuation phenomena. *European Review of Social Psychology*, 6(1), 161–198.
- Rheingold, H. (1993). *The Virtual Community: Homesteading on the Electronic Frontier*. Reading, MA: Addison-Wesley Publishing Co.
- Rosenberg, H., Syed, S., & Rezaie, S. (2020). The Twitter pandemic: The critical role of Twitter in the dissemination of medical information and misinformation during the COVID-19 pandemic. *Canadian Journal of Emergency Medicine*, 22(4), 418–421.
- Rosenkrantz, P., Vogel, S., Bee, H., Broverman, I., & Broverman, D. M. (1968). Sex-role stereotypes and self-concepts in college students. *Journal of Consulting and Clinical Psychology*, 32(3), 287.
- Russell, L. (2020). *Glitch feminism: A manifesto*. Verso Books.
- Scott, C. R., & Rains, S. A. (2020). (Dis) connections in anonymous communication theory: Exploring conceptualizations of anonymity in communication research. *Annals of the International Communication Association*, 44(4), 385–400.
- Short, J., Williams, E., & Christie, B. (1976). *The Social Psychology of Telecommunications*. Hoboken, NJ: John Wiley & Sons, Ltd
- Steiner, P. (1993, July 5) *On the Internet, nobody knows you're a dog [Comic]*. The New Yorker. Retrieved from <http://archives.newyorker.com/?iid=15713&startpage=page0000063#folio=CV1>

- Strandberg, K., & Grönlund, K. (2012). Online deliberation and its outcome—evidence from the virtual polity experiment. *Journal of Information Technology & Politics*, 9(2), 167–184.
- Suler, J. (2001). The psychology of avatars and graphical space in multimedia chat communities. In M. Beißwenger (ed.), *Chat-Kommunikation*. Stuttgart: Ibidem-Verlag.
- Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & Behavior*, 7(3), 321–326.
- Tajfel, H., & Turner, J. C. (2004). The social identity theory of intergroup behavior. In *Political psychology* (pp. 276–293). Psychology Press.
- Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.
- Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., ... & Nyhan, B. (2018). Social media, political polarization, and political disinformation: A review of the scientific literature. Available at SSRN: <https://ssrn.com/abstract=3144139> or <http://dx.doi.org/10.2139/ssrn.3144139>.
- Turkle, S. (2011). *Life on the Screen*. Simon and Schuster.
- Van Der Heide, B., Schumaker, E. M., Peterson, A. M., & Jones, E. B. (2013). The Proteus Effect in dyadic communication: Examining the effect of avatar appearance in computer-mediated dyadic interaction. *Communication Research*, 40(6), 838–860.
- Walther, J. B., Van Der Heide, B., Ramirez Jr, A., Burgoon, J. K., & Peña, J. (2015). Interpersonal and hyperpersonal dimensions of computer-mediated communication. In S. Shyam Sundar (ed.) *The Handbook of the Psychology of Communication Technology*, (pp. 1–22). <https://doi.org/10.1002/9781118426456.ch1>
- Williams, J. E., & Best, D. L. (1990). *Measuring sex stereotypes: A multinational study*. Newbury Park, CA: Sage Publications.
- Williams, K. S. (2005). On-line anonymity, deindividuation and freedom of expression and privacy. *Penn State Law Review*, 110, 687.
- Wohn, D. Y. (2019, May). Volunteer moderators in twitch micro communities: How they get involved, the roles they play, and the emotional labor they experience. In

CHI'19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (Paper 160, 13 pages).
<https://doi.org/10.1145/3290605.3300390>

- Wright, S. (2012). Politics as usual? Revolution, normalization and a new agenda for online deliberation. *New Media & Society*, 14(2), 244–261.
- Wright, S., Graham, T., & Jackson, D. (2016). Third space, social media and everyday political talk. In: Bruns, A, Enli, G, Skogerbø, E, Larsson, AO and Christensen, C, (eds.) *The Routledge Companion to Social Media and Politics*. (pp. 74–88). New York: Routledge.
- Yee, N., & Bailenson, J. (2007). The Proteus Effect: The effect of transformed self-representation on behavior. *Human Communication Research*, 33(3), 271–290.
- Yee, N., & Bailenson, J. N. (2009). The difference between being and seeing: The relative contribution of self-perception and priming to behavioral changes via digital self-representation. *Media Psychology*, 12(2), 195–209.
- Yee, N. (2007). The Proteus Effect: Behavioral modification via transformations of digital self-representation. *Unpublished Doctoral Dissertation*, Stanford University. https://www.nickye.com/pubs/Dissertation_Nick_Yee.pdf
- Yoon, P., & Leem, J. (2021). The Influence of Social Presence in Online Classes Using Virtual Conferencing: Relationships between Group Cohesion, Group Efficacy, and Academic Performance. *Sustainability*, 13(4), 198
- Yun, H. (2006). The creation and validation of a perceived anonymity scale based on the social information processing model and its nomological network test in an online social support community. *Unpublished Doctoral Dissertation*, Michigan State University. <https://doi.org/doi:10.25335/M5G44J28P>
- Zimbardo, P. G . (1969). The human choice: Individuation, reason, and order versus deindividuation, impulse and chaos. In W. J. Arnold & D. Levine (Eds.), *Nebraska Symposium on Motivation*. Lincoln, NB: University of Nebraska Press
- Zimbardo, P. G., Haney, C., Banks, W. C., & Jaffe, D. (1971). *The Stanford prison experiment*. Zimbardo, Incorporated.

Curriculum Vitae

