ANONYMITY, IDENTITY, AND LIES

Artur Pericles L. Monteiro [*]

(from the Knight Institute's Lies, Free Speech, and the Law symposium[**])

Introduction 552

I. The Plurality of Identification: Real Names and Context Collapse 558

A. Before Real Names 559

B. Real Names in Context Collapse 560

II. The Plurality of Anonymity: Norm Conformity and the Mediation of Other Affordances 562

[*] Schmidt Visiting Scholar, International Security Studies, and Lecturer, Yale Jackson School of Global Affairs; Resident Fellow, Information Society Project, Yale Law School; Associate Fellow, Constitution, Law & Politics Research Group, University of São Paulo. I am indebted to Genevieve Lakier, Katy Glenn Bass, Mia Speier, Victoria Tang, and many others with the Knight First Amendment Institute for invaluable editing. This paper draws on a draft presented at Yale Law School's IX Freedom of Expression Scholars Conference (2021), whose participants I thank. I am also grateful to the participants of the Information Society Project writing workshop at Yale Law School, the workshop at the Constitution, Law & Politics research cluster and the Pós-Debate workshop at the University of São Paulo, and the International Security Studies Colloquium at the Yale Jackson School of Global Affairs. As this has been in the works for years, I was fortunate to discuss the paper with many scholars and I have a long list of people whose thoughtful feedback I should acknowledge personally: Aaron Mendon–Plasek, Adriane Sanctis, Amin Afrouzi, Ana Laura P. Barbosa, Arne Westad, Arthur Guerra Filho, Bonnie Kaplan, Camilla Gomes, Charles Fawell, Chinmayi Arun, Christina Spiesel, David Greene, Enrique Armijo, Eric Goldman, Erin Miller, Francesca Procaccini, Goitom Gebreluel, Helen Norton, Ian Seavey, Jisu Kim, Júlia Cani, Kat Geddes, Laura Portuondo, Margot Kaminski, Mattie Webb, Mila Samdub, Michael Karanicolas, Niklas Eder, Nikolas Guggenberger, Nina Santos, Rafael B. Nunes, Rebecca Tushnet, Robert Post, Sean Case, and Virgílio Afonso da Silva.

[**] This Article is being reprinted essentially as it was published on the Knight Institute site, and adheres to the Institute's stylistic preferences. More information about the Lies, Free Speech, and the Law symposium is available at https://knightcolumbia.org/events/symposium-lies-free-speech-and-the-law.

A. Anonymity Does Not Mean Absence of Social Norms 563

B. Affordances and Norms Shape Identity Even in Anonymous Settings 565

C. Measuring the Impact of Anonymity: The Role of Content Moderation 571

III. False Information, Polarization, and Identity 575

A. Identity and Affective Polarization 578

B. Performing Lies and Misinformation: Identification as a Driver 581

C. Anonymity as a Depolarizing, Discourse-Enabling Device 588

Conclusion 593

INTRODUCTION

Anonymity has emerged in recent years as an important focus of debates about the digital public sphere.[1] An opinion piece in The Wall Street Journal argued that a solution to the problems besetting social media was to "end anonymity."[2] Soon after, Senator John Kennedy announced he would introduce a bill to ban anonymity online.[3] In the United Kingdom, anonymity also featured in the discussions about the Online Safety Act.[4] Bills designed to curb anonymity are also frequent in

[1] Olivier Sylvain, *Intermediary Design Duties*, 50 CONN. L. REV. 203 (2017); Danielle Keats Citron & Benjamin Wittes, *The Internet Will Not Break: Denying Bad Samaritans § 230 Immunity*, 86 FORDHAM L. REV. 401 (2017); Mary Anne Franks, *Beyond the Public Square: Imagining Digital Democracy*, YALE L.J. F. 427 (2021).

[2] Andy Kessler, *Online Speech Wars Are Here to Stay*, WALL ST. J. (Jan. 24, 2021), https://www.wsj.com/articles/online-speech-wars-are-here-to-stay-11611526491.

[3] Mike Masnick, *No, Getting Rid of Anonymity Will Not Fix Social Media; It Will Cause More Problems*, TECHDIRT (Feb. 1, 2021), https://www.techdirt.com/articles/20210131/01114246154/no-getting-rid-anonymity-will-not-fix-social-media-it-will-cause-more-problems.shtml.

[4] Online Safety Act 2023. The final version creates a duty for providers falling under the most intense requirements (Category 1 services) to "offer all adult users of the service the option to verify their identity." See *id.*, § 64(1). The Act does not require providers to review official government identification for verification. See § 64(2). While not banning anonymity, the Act also requires providers to offer "features which adult users may use or apply if they wish to filter out non-verified users." § 15(9). There had been calls for a stronger stance against anonymity, but the government decided to adopt a strategy it described as empowering users and striking a balance. *See* Nadine

Brazil, including more recently with one introduced by the select committee investigating the Bolsonaro administration's handling of the pandemic as part of the committee's recommendations included in the final report to punish those who engage in disinformation.[5]

Supporters of proposals targeting anonymity sometimes argue that requiring users to make themselves known will remedy many of the pathologies afflicting the digital public sphere, including misinformation. Identification is seen as a tool for creating a more truth-based discourse, by inducing speakers to behave more responsibly, as well as providing listeners with information to assess the credibility of the speaker. The assumption often is that anonymity promotes lies and incivility, while identification induces truth and civility.[6] Nathaniel Persily sums it up: "If

Dorries, *New Plans to Protect People From Anonymous Trolls Online*, GOV.UK (Feb. 22, 2022), https: //www.gov.uk/government/news/new-plans-to-protect-people-from-anonymous-trolls-online.

[5] *See* Senado Federal, CPI da Pandemia, Parecer No. 1, de 26 de outubro de 2021, 1150 (introducing a requirement that "providers of social networks" verify users' identification including through the use of biometrical data and official taxpayer databases). Under the Brazilian Constitution, "anonymity is forbidden." Constituição Federal [C.F.] [Constitution] art. 5, IV. What that clause entails is unclear and disputed. *See* ARTUR PERICLES LIMA MONTEIRO, ONLINE ANONYMITY IN BRAZIL: IDENTIFICATION AND THE DIGNITY IN WEARING A MASK (2017) (arguing that neither the Constitution nor Brazilian statutory law create a general identification requirement, contrary to what is often stated).

[6] McIntyre v. Ohio Elections Comm'n, 514 U.S. 334, 382 (1995) (Scalia, J., dissenting): ". . . a person who is required to put his name to a document is much less likely to lie than one who can lie anonymously . . . ." *See also* NATHANIEL PERSILY, THE INTERNET'S CHALLENGE TO DEMOCRACY: FRAMING THE PROBLEM AND ASSESSING REFORMS 16 (2020): "When it comes to elections, though, the unaccountable speech anonymity facilitates can promote division and deception that hinders the proper functioning of a democracy. It enables extremist voices that seek to undercut the legitimacy of the electoral process and basic constitutional values. Anonymity and pseudonymity (adopting an online persona other than one's own) also facilitate the kind of lying and misrepresentation that undercut a well-informed electorate."; *See also* Enrique Armijo, *Meet the New Governors, Same as the Old Governors*, *in* THE PERILOUS PUBLIC SQUARE: STRUCTURAL THREATS TO FREE EXPRESSION TODAY 352, 356–57 (David E. Pozen ed., 2019) ("Anonymity, at least as a First Amendment–informed design principle for communications networks, tends to result in a degraded expressive environment, not an improved one."); Anne Wells Branscomb, *Anonymity, Autonomy, and Accountability: Challenges to the First Amendment in Cyberspaces*, 104 YALE L.J. 1639, 1645 (1995) (stating anonymity "strips users of the civility that face-to-face the encounter has engendered in most modern societies" and "facilitates the distribution of false information").

online anonymity is the cause of many of the democracy-related ills of social media, then disclosure might be the best disinfectant."[7]

In fact, in an environment beset by political polarization, instead of serving as a disinfectant, identification can add fuel to the fire of mis- and disinformation. Not only that, anonymity can have a role also in enabling public political deliberation that has been underappreciated. This paper surveys literature from multiple disciplines and challenges assumptions behind the prevailing stances towards anonymity and mis- and disinformation. It argues that anonymity and identification do not have a fixed function;[8] it instead refers to the *plurality of identification and the plurality of anonymity.*[9] "Plurality" is meant to emphasize that both anonymity and identification shape and are shaped by factors such as social norms and platform affordances. As such, whether identification will contribute to a more truth-based public discourse and to a more civic-minded digital sphere is a question that can only be answered if we account for those factors. Considering the identity-based components of the spread of disinformation in polarized contexts, anonymity can serve as a device to create opportunities for conversation and avoid some of the mechanisms triggering those components.

A few notes on terminology and scope should be helpful. Anonymity stands for namelessness in the vernacular, yet conceptually it must be appreciated as going beyond names.[10] In fact, names are less effective as unique identifiers, as they often can shared by more than one person.[11] Identification correspondingly is not constrained to names. Identification and anonymity can be seen as "different poles of

[7] PERSILY, *supra* note 6 at 41.

[8] Because it argues that we must acknowledge that identity is not static, instead of using "identity disclosure" (which seems to imply there is only one identity), this paper prefers the term "identification" to refer to a particular form of identity manifestation. In the context of social media governance and misinformation, this usually refers to adopting real names as an identifier.

[9] I thank Helen Norton for suggesting the phrase "the diversity of anonymity" at Yale Law School's IX Freedom of Expression Scholars Conference. Reflection on that notion prompted this concept.

[10] *See* Helen Nissenbaum, *The Meaning of Anonymity in an Information Age*, 15 INFO. SOC'Y 141, 141 (1999).

[11] *See* Gary T. Marx, *What's in a Name? Some Reflections on the Sociology of Anonymity*, 15 INFO. SOC'Y 99, 101 (1999).

a continuum."[12] Anonymity is relational: Someone might have knowledge that allows them to identify a speaker, while another person might not.[13]

This relational aspect can be relevant particularly when we are considering illegal content, where it is not just listeners who pass judgment on anonymous speech, but also authorities seeking to hold speakers accountable. That is, the audience not having knowledge that identifies a speaker (because their name is not unique or the speaker uses a pen name) can be a different issue than law enforcement and other officials being able to trace the speech.[14] Although the questions are connected, this paper will not discuss traceability.[15] It will focus on identification with one kind of identifier, real names, as a lever that commentators and policymakers have turned to with the aspiration of governing *legal* speech.[16] Combatting

[12] Craig R. Scott & Stephen A. Rains, *(Dis)connections in Anonymous Communication Theory: Exploring Conceptualizations of Anonymity in Communication Research*, 44 ANN. INT'L COMM. ASS'N 1, 392 (2020).

[13] *See* Kathleen A. Wallace, *Anonymity*, 1 ETHICS & INFO. TECH. 23, 24 (1999) ("Anonymity presupposes social relations. In other words, it is relative to social contexts in which one has the capacity to act, affect or be affected by others, or in which the knowledge or lack of knowledge of who a person is is [sic] relevant to their acting, affecting or being affected by others.").

[14] *See* Margot E. Kaminski, *Real Masks and Real Name Policies: Applying Anti-Mask Case Law to Anonymous Online Speech*, 23 FORDHAM INTELL. PROP. MEDIA & ENTM'T L.J. 815, 877 (2013) ("Policies that prohibit anonymity apply to all layers of the communication stack: the individual cannot speak without self-identifying to everyone. Policies that address traceability do not mandate that an individual speak under his real name; instead, they require the individual to register identity with at least one party, so that if he commits a crime or a tort, law enforcement will be able to find him.").

[15] For a discussion of untraceable anonymity, *see* A. Michael Froomkin, *From Anonymity to Identification*, 1 J. SELF-REGULATION & REG. 120 (2015).

[16] This was one argument in Justice Scalia's dissent in a leading precedent protecting anonymity. He noted that identification played a part in "promoting a civil and dignified level of campaign debate—which the State has no power to command, but ample power to encourage by such undemanding measures as a signature requirement [for campaign material]." 514 U.S. 334, 382 (1995) (Scalia, J., dissenting).

mis- and disinformation is one reason why commentators want to expand identification. One shared hope is that both speakers and listeners will be closer to the truth through real-name identification.[17] That is the central concern of this paper.

Part I introduces the concept of the plurality of identification, which the paper uses to call attention to how real names have a different operation on social media. Names, which were not ubiquitously employed to the same extent they are now (e.g., full names in Facebook profiles), work in markedly transformed ways when they offer an index to massively aggregated, permanent information on every one of us that is accessible through social media and search engines. Calls for identification often rest on an assumption that real names instantiate the same identity regardless of the context they are displayed. This ignores the impact of context collapse[18]—the flattening of different social contexts—in impelling individuals to perform their identity to an imagined, unspecified audience with which they engage much like micro-celebrities.

At the same time, anonymity is thought to prevent accountability by disconnecting us from drivers of norm-abiding behavior. Part II shows that this is only sometimes true—and introduces the plurality of anonymity. It surveys research establishing that anonymous settings may produce *greater* conformity to local, i.e., group-related, social norms (which may or may not be democratically desirable). The paper then argues that the impact of anonymity on user behavior depends on content moderation practices and community norms.

Part III consolidates those points and discusses the role of political polarization and the sharing of false information. Although it is commonly assumed that identification is a means of fostering veracity, as well as civility, this is often not the case. The paper explores findings from psychology and computational social science to argue that real names are part of mechanisms that drive misinformation in settings marked by affective polarization (negative attitudes toward the other party). Anonymity, conversely, has potential as a device for reducing polarization as well as creating opportunities for conversations not infected by those mechanisms.

[17] Seth Kreimer refers to this as "purification by publicity." Seth F. Kreimer, *Sunlight, Secrets and Scarlet Letters: The Tension Between Privacy and Disclosure in Constitutional Law*, 140 U. PA. L. REV. 1, 89 (1991). See also Part III, *infra*, notes 104-111 and accompanying text.

[18] *See* DANAH BOYD, *Social Network Sites as Networked Publics: Affordances, Dynamics, and Implications*, *in* A NETWORKED SELF: IDENTITY, COMMUNITY, AND CULTURE ON SOCIAL NETWORK SITES 39, 49 (Zizi Papacharissi ed., 2011) (describing context collapse).

This paper aims to add to a years-long debate about the place of anonymity in a healthy digital public sphere.[19] Much work has been done about the disproportionate effects flowing from real-name policies to marginalized communities, to individuals who have legitimate reason to fear for their safety in disclosing their real names, or to those whose names do not match their official government identification. Indeed, in 2011, the announcement of now-defunct Google Plus's real-name policy prompted considerable backlash along those lines, leading in what were described as the Nymwars;[20] in 2015, a new battlefront turned to changes in Facebook's enforcement of its policies, which was met with opposition by a collection of civil society organizations gathered around the Nameless Coalition.[21] Scholars have suggested that such concerns can be addressed in specific cases and exceptionally, only "where anonymity is needed to avoid 'threats, harassment, or reprisals,'" as Justice Scalia argued in *McIntyre*, a landmark case on the topic.[22] My hope with this paper is to explore the role of anonymity and identification even beyond the risk of speech suppression and disproportionate effects.

[19] For legal scholarship, see JEFF KOSSEFF, THE UNITED STATES OF ANONYMOUS: HOW THE FIRST AMENDMENT SHAPED ONLINE SPEECH (2022) for a recent overview. For a theoretical discussion that also explores anonymity regulation beyond the U.S., see ERIC BARENDT, ANONYMOUS SPEECH: LITERATURE, LAW AND POLITICS (2016). See also, among others, Branscomb, *supra* note 6; Lee Tien, Who's Afraid of Anonymous Speech? *McIntyre* and the Internet, 75 OR. L. REV. 117 (1996); A. Michael Froomkin, *Legal Issues in Anonymity and Pseudonymity*, 15 INFO. SOC'Y 113 (1999); Danielle Keats Citron, *Cyber Civil Rights,* 89 B. U. L. REV. 61 (2009); Lyrissa Barnett Lidsky & Thomas F Cotter. *Authorship, Audiences, and Anonymous Speech*, 82 *Notre Dame L. Rev.* 1537 (2007); Kaminski *supra* note 14; Rebecca Tushnet, *The Yes Men and The Women Men Don't See*, *in A WORLD WITHOUT PRIVACY* (Austin Sarat ed., 2014). A. Michael Froomkin, *Lessons Learned Too Well: Anonymity in a Time of Surveillance*, 59 ARIZ. L. REV. 95 (2017).

[20] *See* Eva Galperin, *2011 in review: Nymwars*, DEEPLINKS (2011), https://www.eff.org/deeplinks/2011/12/2011-review-nymwars. *See also* Jillian C. York, *A case for pseudonyms*, DEEPLINKS (2011), https://www.eff.org/deeplinks/2011/07/case-pseudonyms.

[21] *See* Eva Galperin & Wafa Ben Hassine, *Changes to Facebook's "real names" policy still don't fix the problem*, DEEPLINKS (2015), https://www.eff.org/deeplinks/2015/12/changes-facebooks-real-names-policy-still-dont-fix-problem.

[22] 514 U.S. 334, 385 (1995) (Scalia, J., dissenting) (citing NAACP v. Alabama ex rel. Patterson, 357 U. S. 449 (1958)). *See* BARENDT, *supra* note 19 at 68–70, 80 (arguing that anonymity should only be protected in circumstances where "its value clearly outweighs the risks"); Helen Norton, *Secrets, Lies, and Disclosure*, 27 J.L. & POL. 641, 646 (2012) (urging "that we add an inquiry into why speakers want to keep their identity secret to the factors that we consider when thinking about disclosure requirements' First Amendment autonomy implications.").

I. The Plurality of Identification: Real Names and Context Collapse

Mark Zuckerberg framed real names as the appropriate norm for online interaction when he claimed, in 2010, that "[h]aving two identities for yourself is an example of a lack of integrity."[23] The notion seems to be: People hardly ever use assumed names in offline life, so why should they do it differently online?[24]

By implementing a real-name policy for Facebook, and its 3 billion users worldwide, Zuckerberg gave some credit to the notion that using real names is what is to be expected generally from people online. He made his assertion a kind of self-fulfilling prophecy.

Discussions of online anonymity often frame it as a *deviation* from established social norms[25]—a deviation that is justified by an individual's legitimate fear of retaliation,[26] or as a legitimate response to surveillance.[27] However, despite the allure of the familiarity of names in a pre-internet, offline world, it is instead real-name policies that break with longstanding conventions. As Section A shows, the internet did not always presume real names.

Even when users adopt real names online, doing so significantly alters the function those names play, as Section B explores. This is because those names get indexed on multiple social media and search engines, the result being that users' multiple audiences, representing a range of social interactions, are flattened into a single one.[28] This forces them to perform to their "most sensitive [audience] members: parents, partners, and bosses,"[29] as if they were broadcasting for these networked

[23] BERNIE HOGAN, *Pseudonyms and the Rise of the Real-Name Web*, *in* A COMPANION TO NEW MEDIA DYNAMICS 290, 292 (John Hartley, Jean Burgess, & Axel Bruns eds., 2013).

[24] danah boyd, *The Politics of "Real Names,"* 55 COMM. ACM 29, 30 (2012) (arguing that "real name" policies rely on an implicit notion of the role of names in offline interactions). For an analysis of the role of "real names" in the early days of Facebook and of statements by Mark Zuckerberg on authenticity, see Oliver L. Haimson & Anna Lauren Hoffmann, *Constructing and Enforcing "Authentic" Identity Online: Facebook, Real Names, and Non-Normative Identities*, 21 FIRST MONDAY 6 (2016).

[25] boyd, *supra* note 24 at 30 (discussing how names shift "how people relate online").

[26] PAUL BERNAL, THE INTERNET, WARTS AND ALL 220–23 (2018).

[27] Froomkin, *supra* note 15.

[28] See *infra* Part II, Section I.B.

[29] Alice E. Marwick & danah boyd, *I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience*, 13 NEW MEDIA & SOC'Y 114, 125 (2011).

audiences.[30] Real-name identification in such a setting does not mean the same as it would in each social context; this evidences how identification is multifarious.

A. Before Real Names

In fact, the early days of computers and the internet were marked by identifiers other than real names. As Emily van der Nagel reports,[31] the earliest usernames were actually numbers: System administrators would assign individuals unique user identification numbers to distinguish their activities from those of others who shared the same computer (then owned only by institutions).

And although early email accounts were at first controlled by institutions, not individuals—who tended to use employees' or students' full names as their email identification (or a combination of initials and numbers)[32]— as the internet developed and the commercial internet grew, service providers started offering personal email addresses for a fee. Once (institutional and financial) constraints on email creation disappeared, people began to choose usernames creatively, "play[ing] with numbers, nicknames, interests, in-jokes and cultural references."[33] Those creative email addresses were also a way to establish boundaries between work and personal life, which made more sense before connected portable devices (laptops and phones) eroded those divisions.

Pseudonyms were also a staple of users of social media precursors such as bulletin boards and IRC (internet relay chat) channels.[34] As early as the 1970s, people played with usernames at the Electronic Information Exchange System, a computer conferencing bulletin board, so they could adopt "a role in particular conferences, have the freedom to say things they would not want attributed to them or their organisation, signal that the discussion was not to be taken too seriously, and let newcomers experiment with sending messages on the board without fear of revealing their lack of skill in the medium." Foundational work by Sherry Turkle, writing on

[30] *Id.* at 129.

[31] Emily van der Nagel, *From Usernames to Profiles: The Development of Pseudonymity in Internet Communication*, 1 INTERNET HISTORIES 312 (2017).

[32] *Id.* at 315.

[33] *Id.* at 316.

[34] "Since 1988, IRC has allowed people to exchange text-based messages in dedicated channels modelled after citizens' band radio, first within Finland, then across the global Internet." *Id.* at 319. See also *id.* at 320–22 for user interface illustrations.

the early days of the commercial internet, discussed how users in IRC channels had fluid identities and explored how this helped to create a space in which conventions around gender, age, and race could be redefined and transformed.[35] Those hopes did not bear out as Turkle might have expected, in large part because "forms of discrimination such as racism and sexism are not solely based on appearance."[36]

The tendency of users to continue to rely on pseudonyms was a consequence of many features of the early internet, which Bernie Hogan discusses.[37] First, pre-Web 2.0, user-generated content was generally text-based, and digital cameras and webcams were not yet widespread. As such, constructing a new identity required less effort. Second, because relatively few people used the internet, communities tended to be interest-based, not based on social ties. This meant there were few costs for using pseudonyms online. Third, the internet was still a mystery to many, even those who used it, and people were wary of exposing their "real-world" identities.

B. Real Names in Context Collapse

The rise of social media altered many of these features of the internet. And because social media platforms were designed to link people to those they were already connected to offline in some way, it made sense for users to employ their real names when they used the platforms.[38] Indeed, it is worth remembering that Facebook was early on described as an online version of Harvard's paper face books.[39] Real names made sense for TheFacebook, just as pseudonyms made sense for other websites. Initially, Facebook was limited to the Harvard community; it would later be extended to other universities in the US. Still, it was a walled garden, with social norms appropriate for the context of that community.

[35] SHERRY TURKLE, LIFE ON THE SCREEN: IDENTITY IN THE AGE OF THE INTERNET (1995).

[36] ALICE E. MARWICK, *Online Identity*, *in* A COMPANION TO NEW MEDIA DYNAMICS 355, 357–58 (John Hartley, Jean Burgess, & Axel Bruns eds., 2013).

[37] *See* HOGAN, *supra* note 23 at 292.

[38] boyd, *supra* note 24 at 29–30.

[39] Alan J. Tabak, *Hundreds Register for New Facebook Website*, HARV. CRIMSON (Feb. 9, 2004), https://www.thecrimson.com/article/2004/2/9/hundreds-register-for-new-facebook-website/.

An important change happened, however, when the platform became accessible to anyone with an email.[40] Users now could interact simultaneously with their high school friends, college colleagues, family, coworkers, and so on. This meant that users had no single set of social norms they could rely upon when communicating to these multiple audiences. The opening up of Facebook resulted, in other words, in context collapse, a term which stands for "[t]he lack of spatial, social, and temporal boundaries mak[ing] it difficult to maintain distinct social contexts."[41]

The consequence was that, although Facebook's real-name policy stuck around, users' real names no longer played the role they did offline. For one thing, users' real names were now persistent and searchable: When users spoke online, their words were not only broadcast to everyone in their online network; they could be found and associated with them at any later point. With context collapse, users would be read by audiences they might not have expected. Attempting to make a joke after giving the barista one's name entailed the risk of either looking silly to a handful of people nearby or drawing a few chuckles from them. With real-name social media accounts, the embarrassment goes much further, as does the comedy. This is not just a question of reach; it affects how users see themselves and what they post.

Alice Marwick and danah boyd have explored this transformation in how people interact online. They show how the collapsed social contexts drive people to engage in practices of "micro-celebrities," much like broadcast television, with the caveat that "unlike broadcast television, social media users are not professional image-makers."[42] To the extent that each social interaction enacts identity, the collapsing of contexts in social media means that users must present themselves to an imagined audience (who they think might consume their content) that does not share a set of norms regarding what is appropriate.

[40] *See* boyd, *supra* note 24 at 29: "At Harvard, Facebook's launch signaled a safe, intimate alternative to the popular social network sites. People provided their names because they saw the site as an extension of campus life. . . . As Facebook spread beyond college campuses, not all new users embraced the 'real names' norm. During the course of my research, I found that late teen adopters were far less likely to use their given name. Yet, although Facebook required compliance, it tended not to actively—or at least, publicly—enforce its policy."

[41] BOYD, *supra* note 18 at 49.

[42] Marwick & boyd, *supra* note 29 at 123.

So, while sticking to real names might seem a continuation of established social practices, it is not, because internet affordances change how names operate socially. Our names are indexed, our café encounters, our workplace banter, our relationships—in short, now we are visible to all, we have to perform our identities for all those people, or pay the price for not doing so.

In light of that, we can see that pseudonyms in fact make sense online, because they allow people to navigate different contexts, and speak in different registers to different audiences.[43] This is not to say that real names on social media do not make sense. Billions of users found value in connecting to high school friends, distant family members, former coworkers, etc. The point that we should be clear on is how real names online are not a continuation of our pre-digital practices. And, as Part I, Section A showed, the ensuing transformation is not directly a result of technological change. As Bernie Hogan notes, "[t]he real-name web is not a technology; it is a practice and a system of values."[44] The familiar appeal of using real names, therefore, rests on an inadequate understanding of how internet affordances changed what our names mean. The impact of attaching real names to our speech and actions varies, and this is how we can see the plurality of identification.

II. THE PLURALITY OF ANONYMITY: NORM CONFORMITY AND THE MEDIATION OF OTHER AFFORDANCES

Part I explored how the same form of identification can function differently according to the context. Real names have different implications in digital settings. Part III will explore how this variation frustrates the assumptions of commentators who put faith in identification to combat mis- and disinformation. This Part shows how anonymity can play a part in making behavior conform to social norms, a point that is often neglected. Section A introduces the theoretical model that describes how. Section B then transitions from theory to practice. It canvasses some of the ways anonymous communities work to shape identities around their aspirations and goals. Section C discusses quantitative research that has sought to understand the role of anonymity in the quality of online content by studying newspaper comment sections.

[43] *See* Tushnet, *supra* note 19 at 86–89.

[44] HOGAN, *supra* note 23 at 291.

A. Anonymity Does Not Mean Absence of Social Norms

It is tempting to think of online anonymity as bringing out the worst in us. If users are not held accountable for their offline identities, the argument goes, then incentives to refrain from engaging in abusive behavior are removed, and only incentives to indulge in toxic disinhibition remain. In short, the idea is that when individuals are anonymous, they will flout social norms and behave badly. This tracks classic theories on deindividuation in social psychology.[45] This familiar view of the impact of anonymity has been challenged in recent decades by scholars in social psychology and communication studies who have developed the social identity model of deindividuation effects (SIDE).[46]

This model holds that, in many situations, "group immersion and anonymity le[a]d to greater *conformity* to specific (i.e., local) group norms, rather than to transgression of general prosocial norms, as deindividuation theory proposed."[47] Contrary to classic deindividuation theory, which links the lack of identification with individuals acting in disdain for any social norms, the SIDE model predicts that, when group identity is salient,[48] it will modulate anonymous individuals.[49]

[45] *See* Felipe Vilanova et al., *Deindividuation: From Le Bon to the Social Identity Model of Deindividuation Effects*, 4 COGENT PSYCHOL. 1308104 (2016).

[46] S. D. Reicher, R. Spears, & T. Postmes, *A Social Identity Model of Deindividuation Phenomena*, 6 EUR. R. SOC. PSYCHOL. 161 (1995).

[47] RUSSELL SPEARS, SOCIAL IDENTITY MODEL OF DEINDIVIDUATION EFFECTS 1, 2 (Patrick Rössler, Cynthia A. Hoffner, & Liesbet van Zoonen eds., 2017).

[48] In experiments, group identity salience is often achieved by manipulating the cues available to participants, through design choices that represent them in terms of the identity researchers are trying to emphasize. This can be done, for instance, by user interfaces that provide only cues to make the group identity salient, instead of photos or names that would give participants information about each other. Identity salience is also manipulated by telling participants that they were selected because they share the same characteristics as other group members—because they are science students, as opposed to social science students (and vice-versa), in one experiment. *See* Reicher, Spears, & Postmes, *supra* note 46 at 177–78 (discussing strategies for operationalizing identity salience).

[49] For previous legal writing discussing deindividuation, see Katherine S. Williams, *On-Line Anonymity, Deindividuation and Freedom of Expression and Privacy*, 110 PENN ST. L. REV. 687, 693–97 (2006); Diane Rowland, *Griping, Bitching and Speaking Your Mind: Defamation and Free Expression on the Internet*, 110 PENN ST. L. REV. 519, 531–35 (2006); Julie Seaman, *Hate Speech and Identity Politics: A Situationalist Proposal*, 36 FLA. ST. U. L. REV. 99, 116–21 (2019).

Deindividuation theory would see the behavior of individuals in a crowd as irrational and anti-normative, reflecting a sense of loss of identity and the constraints of self-awareness. The SIDE model sees such behavior as a consequence of individuals corresponding to group identity and local norms, acting in accordance with what that group finds normative. In a nutshell, where deindividuation theory "implies a loss of self in the group," the SIDE model instead recognizes "the emergence of the group in the self" [50]—when individuals perceive each other as "interchangeable group members."[51] Initially applied to text-based media, the model has been extended to other kinds of media as well (e.g., video-based).[52] The SIDE model has been supported by multiple research findings.[53]

So the notion that online anonymity entails a negation of identity and any kind of social norms must be revised in light of research showing how, even in conditions of anonymity, identities are still intermediated by norms. We should be careful about what this means. It does not mean that group identity and corresponding norms will always prevail. Which identity will be salient depends on a wide range of factors; the SIDE model does not say it will always be the case, instead, it rejects

[50] RUSSELL SPEARS & TOM POSTMES, *Group Identity, Social Influence, and Collective Action Online: Extensions and Applications of the SIDE Model*, *in* THE HANDBOOK OF THE PSYCHOLOGY OF COMMUNICATION TECHNOLOGY 23, 27 (S. Shyam Sundar ed., 2015). Somewhat confusingly, despite the name "social identity model of *deindividuation* effects," proponents of the SIDE model reject the notion of a deindividuated state (how deindividuation theory describes the lack of social regulation). *See* Spears and Postmes, *id.* at 29–30. Instead, they refer to the process through which group identity governs as "depersonalization." Note, however, that the term does not imply that individuals are then not acting as persons, but instead that their actions are better explained by the *impersonal* perspective of the group.

[51] SPEARS, *supra* note 47 at 3.

[52] See SPEARS & POSTMES, *supra* note 50 at 34–36 (discussing research beyond text-based media).

[53] *See* SPEARS & POSTMES, *supra* note 50 (reviewing evidence supporting the model). *See also* Guanxiong Huang & Kang Li, *The Effect of Anonymity on Conformity to Group Norms in Online Contexts: A Meta-Analysis*, 10 INT'L J. COMM. 398, 16 (2016) (meta-analysis reviewing 13 studies, concluding that "[The] result supports the SIDE model, such that anonymous individuals define their identities on a group level, and their behaviors are guided by the norms associated with their salient group memberships."). This is not to say that the SIDE model has been definitively proven as true and that deindividuation theory has been abandoned. *See* Vilanova et al., *supra* note 45 (arguing the SIDE model does not replace but actually supplements deindividuation).

a "blanket assumption that people will always act in line with *individual* self-interest when anonymous."[54] It also does not mean that the resulting norms will guide group behavior toward positive social outcomes. Importantly, the norms here are local, i.e., those embraced by the group, and might be in tension with broader social norms or with the law.

Indeed, as noted, SIDE explains (instead of refuting or ignoring) how, in groups such as mobs, individuals can be guided toward extreme conduct. While one might think that the anonymity of the mob (i.e., the fact that individual behavior is less likely to be discerned) releases mob members from social norms, the reverse is often the case: Individuals are dragged by the mass behavior because they fused to the (destructive) group identity. The insight borne out by this framework is that this is not a result of the *absence* of social norms. It is rather the opposite: Groups can become more extreme than the aggregation of members' attitudes precisely *because* group identity plays such an overwhelming force.[55] We turn now to 4chan and Reddit to see in practice how group identity can be shaped to very different results.

B. Affordances and Norms Shape Identity Even in Anonymous Settings

The SIDE model shows that we should not assume that anonymity necessarily erodes the constraints of identity and social norms. Identity can play a part in anonymous settings, and identity performance is then not unlike what takes place in non-anonymous settings when we perform not just one but many roles (or, under context collapse, try to negotiate performing those identities for audiences with differing expectations). How we make decisions regarding identity performance in such circumstances is the result of the interplay of digital affordances and social norms, which are reciprocally shaped.

[54] SPEARS & POSTMES, *supra* note 50 at 32.

[55] *See id.* at 25, for a brief overview of how social identity theory, on which the SIDE model builds, explains the "group polarization, in which group discussion results in group decisions that are more extreme (or 'polarized') than the mathematical average of individual group members' attitudes."

The outcome of this complicated function can affirm or undermine our democratic aspirations for the digital public sphere. The argument here is not that anonymity always yields valuable results. Instead, it is that the role of anonymity in that function is not linearly fixed. Commentators often talk as if it were.[56]

To see how, we can consider platforms that allow users to be anonymous and where anonymity is the norm—and are still markedly different. 4chan and Reddit both enable users to post without any verification.[57] Users can employ multiple handles and create temporary accounts (which on Reddit are known as throwaway accounts),[58] one for each post they want to make even;[59] 4chan goes a step further and allows for the same handle to be shared by multiple users, which is the norm.[60] They fall roughly on the same extreme of the spectrum from real-name verified accounts and no identification at all. In spite of that, the 4chan boards and Reddit subreddits that we will consider are starkly contrasting.

[56] For Kyle Langvardt, "[t]he possibility of anonymous speech on the Internet, combined with the ease of 'one to many' communications, largely removes the normative and practical constraints that made content-shock rare in the twentieth century." Kyle Langvardt, *Regulating Online Content Moderation*, 106 GEO. L.J. 1353, 1361 (2018); Saul Levmore argues "that one cost of Internet anonymity is that a successful site must monitor and censor in order to inhibit what might become overwhelming noise." SAUL LEVMORE, *The Internet's Anonymity Problem*, *in* THE OFFENSIVE INTERNET 50, 59 (Saul Levmore & Martha Nussbaum eds., 2010); Mary Anne Franks mentions anonymity as part of "many characteristics of virtual interactions [that] negatively impact communication and debate." Franks, *supra* note 1 at 436.

[57] On 4chan, "[t]here is no registration process or login required." Lee Knuttila, *User Unknown: 4chan, Anonymity and Contingency*, 16 FIRST MONDAY 10 (2011); Reddit "allows one-time use accounts that are easily created by signing up only with a new username, password, and CAPTCHA (even an email address is not required)." Alex Leavitt, *"This Is a Throwaway Account": Temporary Technical Identities and Perceptions of Anonymity in a Massive Online Community*, PROC. 18TH ACM CONF. ON COMPUT. SUPPORTED COOP. WORK & SOC. COMPUTING 317, 320 (2015).

[58] *See* Leavitt, *supra* note 57.

[59] *See* Munmun De Choudhury & Sushovan De, *Mental Health Discourse on Reddit: Self-Disclosure, Social Support, and Anonymity*, 8 PROC. INT'L AAAI CONF. WEB & SOC. MEDIA 71, 78 (2014) (2014 study on Reddit finding 61% of the users in a selection of mental health subreddits had posted a single post or comment).

[60] *See* Knuttila, *supra* note 57 ("the vast majority of posts fall under the default username: Anonymous.").

Reddit operates with federated community standards and moderation, with site-wide (or federal) policies and practices supplemented by more specific, community-built and enforced, (local) subreddit rules.[61] Site-wide policies and their enforcement were significantly stiffened after very visible incidents, particularly the use of the website for the non-consensual sharing of intimate images of celebrities, leading the platform to ban a community that had hosted much of the material.[62] After 2015, Reddit announced an update to its harassment policy that culminated in the banning of "a fatphobic community [targeting] photographs and videos of overweight and/or obese persons."[63] Other subreddits were later banned, and the platform also started using quarantine as an enforcement instrument.[64]

Once again, this federal level of policies and enforcement sits on top of communities', which can abide by stringent rules for eligibility to participate (sometimes by obtaining assurances of who the user is without checking any official or

[61] *See* Shagun Jhaver et al., *"Did You Suspect the Post Would Be Removed?": Understanding User Reactions to Content Removals on Reddit*, 3 PROC. ACM HUM.–COMPUT. INTERACTION 1, 5 (2019) ("First, there exists a user agreement and content policy similar to the terms and conditions of many websites. Second, a set of established rules defined by Reddit users, called [r]ediquette, guide site-wide behavior. Finally, many subreddits also have their own set of rules that exist alongside site-wide policy and lay out expectations about content posted on the community."). *See also* Casey Fiesler et al., *Reddit Rules! Characterizing an Ecosystem of Governance*, PROC. 12TH INT'L AAAI CONF. ON WEB & SOC. MEDIA 72 (2018); Eshwar Chandrasekharan et al., *The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales*, 2 PROC. ACM HUM.–COMPUT. INTERACTION 1, 2 (2018).

[62] *See* Julia R. DeCook, *R/WatchRedditDie and the Politics of Reddit's Bans and Quarantines*, 6 INTERNET HISTORIES 206, 212–13 (2022); Adrienne Massanari, *#Gamergate and The Fappening: How Reddit's Algorithm, Governance, and Culture Support Toxic Technocultures*, 19 NEW MEDIA & SOC'Y 329 (2017).

[63] DeCook, *supra* note 62 at 212.

[64] See *id.* at 212 ("[The August 2015] policy update also introduced the quarantine function of [R]eddit, where subreddits are not removed but are kept from reaching the front page and require a user to agree to view the content (effectively creating more friction to access the community).").

institutional forms of identification)[65] and the manner of participation.[66] That shows that group identity is deliberately and fastidiously molded by the communities, promulgating and patrolling the model of behavior they have elected for themselves.[67] That effort by communities sits within Reddit's 'karma' system and upvote and downvote mechanisms,[68] which affects content visibility,[69] and which subreddits can to an extent wield as part of their governance strategies (e.g., by instructing

[65] *See* EMILY VAN DER NAGEL, *Embodied Verification: Linking Identities and Bodies on NSFW Reddit*, *in* MEDIATED INTERFACES: THE BODY ON SOCIAL MEDIA 47, 58 (2020) (describing verification on sexual exhibitionist subreddits "as an act that proves consent, as including a Reddit username, the date, and the name of the subreddit in a photograph with their body is a way of asserting the person posting their selfie took it with the intention of uploading it to Gonewild"); Emily van der Nagel, *Faceless Bodies: Negotiating Technological and Cultura Codes on Reddit Gonewild*, 10 SCAN – J. MEDIA ARTS CULTURE (2013) (same); Tawfiq Ammari, Sarita Schoenebeck, & Daniel Romero, *Self-Declared Throwaway Accounts on Reddit*, 3 PROC. ACM HUM.–COMPUT. INTERACTION 1, 23–24 (2019) (noting that a subreddit for parents asks that those interested in joining provide a link to a post on Reddit corroborating the user has children as well as a picture displaying the handle of the user next to "items only new fathers would have," such as a stroller or diapers).

[66] Subreddit rules often govern not just what the community is about (generally a topic or interest), but also how users should engage. *See* Fiesler et al., *supra* note 61 (describing subreddit rules on formatting posts, links and outside content, off-topic content, low-quality content and others).

[67] A mixed-methods large-scale study on subreddit rules found that subreddits commonly have rules that seek to model personality that is welcome (or unwelcome) in that community. *See id.* at 77 (Table 2, reporting 40.15% of the manually coded, qualitative sample of subreddits that had rules included rules on personality, as did 30.39% of the classifier-based, large-scale data set analysis). Tawfiq Ammari, Sarita Schoenebeck, and Daniel M. Romero, who have researched throwaway accounts used in parenting communities, "argue that throwaways provide parents with shared norms and expectations for sharing potentially stigmatizing experiences while still being embedded within their existing online community." Ammari, Schoenebeck, & Romero, *supra* note 65 at 3.

[68] *See* Tim Squirrell, *Platform Dialectics: The Relationships Between Volunteer Moderators and End Users on Reddit*, 21 NEW MEDIA & SOC'Y 1910, 1922: (2019) "the karma system . . . allows users to 'vote' on content (including 'submissions' – links, images, videos and text posts – and comments on these submissions) and influence its visibility to others. The net 'score' ('upvotes' minus 'downvotes') is displayed next to content, and a user's overall karma from all their submissions and comments is displayed on their (relatively minimal) profile."

[69] *See* Sarah A. Gilbert, *"I Run the World's Largest Historical Outreach Project and It's on a Cesspool of a Website." Moderating a Public Scholarship Site on Reddit: A Case Study of r/AskHistorians*, 4 PROC. ACM HUM.–COMPUT. INTERACTION 1, 4 (2020) ("The total number of votes, or karma, is used to determine what content is seen; although the exact algorithm that determines which posts will be promoted to users' front page or r/all is proprietary, content that is highly

users to use the downvote function to enforce community rules, not so much to signal their disliking of the content).[70] There are also other platform affordances that subreddits can use and adjust to their needs, including automation tools for moderation[71] and "flairs," color tags that can be attached by moderators to both pieces of content and usernames (when displayed in that community). For instance, r/AskHistorians uses flairs as badges of community-verified expertise.[72]

4chan, on the other hand, is decidedly not invested in that kind of meticulously manicured public forum. 4chan message boards such as /b/ are often described as "a well-known trolling stomping ground,"[73] notoriously often accorded the distinction of being one of "the dark corners of the internet."[74] Like Gab and 8chan, 4chan "engage[s] in little or no moderation of the content posted."[75]

That might be taken to suggest that group identity and social norms do not play a role. Yet the opposite is true. Meaningful participation in such 4chan boards in

upvoted rises to the top, while highly downvoted content is obfuscated."). *See also* Squirrell, *supra* note 68 at 1922 ("The consequences of being downvoted are that users are less likely to accord a post credence, while also creating a 'bandwagon' effect, where more users pile in to downvote a post further. Worse, the comment will become invisible to many users: a user-adjustable setting hides posts below a certain point threshold until they are clicked upon.").

[70] *See* Squirrell, *supra* note 68 at 1922–23 (describing how two subreddits dedicated to self-improvement leverage this, and noting communities are constrained by platform-wide affordances and design choices).

[71] *See* Shagun Jhaver et al., *Human–Machine Collaboration for Content Regulation: The Case of Reddit Automoderator*, 26 ACM TRANSACTIONS ON COMPUTER-HUMAN INTERACTION (TOCHI) 1 (2019); Lucas Wright, *Automated Platform Governance Through Visibility and Scale: On the Transformational Power of Automoderator*, 8 SOC. MEDIA + SOC'Y 205630512210770 (2022).

[72] *See* Gilbert, *supra* note 69 at 6 ("A key feature of r/AskHistorians is its panel of experts. The panel system was established so that users could identify experts through the use of flair, a coloured line of text adjacent to the username. Those who want flair must provide evidence of their expertise by linking comments made in r/AskHistorians that demonstrate this expertise. Moderators review these submissions and either award flair or provide feedback on how a submission for flair could be improved.").

[73] DANIELLE CITRON, HATE CRIMES IN CYBERSPACE 53 (2014).

[74] PERSILY, *supra* note 6 at 21.

[75] Richard Ashby Wilson & Molly K. Land, *Hate Speech on Social Media: Content Moderation in Context*, 52 CONN. L. REV. 1029, 1046 (2021).

fact requires intricate demonstrations of membership, which are designed to cordon off outsiders.[76] These range from the digital equivalent of shibboleths (for instance, being able to post unusual Unicode characters),[77] to particular slang,[78] to a choreography involving sarcastic use of design features (such as "memeflags"),[79] grasp of community tropes regarding current affairs, and textual and nontextual representations.[80] Seasoned users explicitly tell the uninitiated to observe and assimilate the ways of the community.[81] Mastery of social norms is persistently tested, and lack of familiarity prompts chastisement. Archetypes about members and unwanted participants are also upheld.[82]

[76] "To communicate high status in the community, most users tend to turn to textual, linguistic, and visual cues." Michael Bernstein et al., *4chan and /b/: An Analysis of Anonymity and Ephemerality in a Large Online Community*, 5 PROC. INT'L AAAI CONF. WEB & SOC. MEDIA 50, 56 (2011).

[77] "One example status signal in /b/ is the classic barrier for newcomers called 'triforcing.' Triforcing means leaving a post using Unicode to mimic the three-triangle icon of popular video game The Legend of Zelda . . . Uninitiated users will then copy and paste an existing triforce into their reply. It will look like a correct triforce in the reply field; however, after posting, the alignment is wrong." *Id.*

[78] "Small gags morph into cultural punch lines and simple misspellings become new popular slang." Knuttila, *supra* note 57. "Simply writing in 4chan dialect is non-obvious to outsiders and in dialect writing serves as an entry-level signal of membership and status." Bernstein et al., *supra* note 76 at 56.

[79] "[E]ach post has a flag included to it. The default flag will indicate the country to which their IP address is located, labeled as 'geographic location.' The other identifiers present are randomly generated thread IDs, which are created for an individual user and persists only within a single thread. Users, however, have other flags available to them, chosen via drop-down menu before the reply is submitted. These alternative options are known as 'memeflags,' and represent ideologies and organizations such as LGBT (lesbian, gay, bisexual, and transgender), the United Nations, Nazi, and others." Dillon Ludemann, *Digital Semaphore: Political Discourse and Identity Negotiation Through 4chan's /Pol/*, NEW MEDIA & SOC'Y 2274, 2729 (2021).

[80] *See id.* (investigating how users demonstrate membership in the 4chan message board /pol/).

[81] "Lack of fluency is dismissed with the phrase 'LURK MOAR,' asking the poster to spend more time learning about the culture of the board." Bernstein et al., *supra* note 76 at 56.

[82] "This was identified by other users as a 'shill' post. In brief, a 'shill' in this context is a person who pretends to lean into a conspiracy to absurdity, often with the intention of discrediting the theory or to deter others from participating and can be considered as trolling to an extent. It is also the assumption that shills are being paid to post, and are frequently met with contempt by others, wherein isolating shills here and trolling them back have become political participation." Ludemann, *supra* note 79 at 2735.

More specifically, the import from the SIDE model is that we should not assume that anonymity works the same on platforms such as 4chan and Reddit. The former does virtually no moderation; community norms are uncodified, and there is often apparent informal approval of abuse and harm toward out-group users. The latter platform, in contrast, operates with federated community standards and moderation, with site-wide (or federal) practices supplemented by more specific, community-built and enforced, (local) subreddit rules. In terms of requiring and validating information, they might otherwise be seen as quite similar. Yet the differences are striking. While certain 4chan boards are often referred to as one of "the dark corners of the internet,"[83] researchers have shown how subreddits are able to create vibrant forums for scholarly knowledge, parenting, and intimate content,[84] among others. The SIDE model offers insight as to why: anonymity is employed with patently different goals— and outcomes.

C. Measuring the Impact of Anonymity: The Role of Content Moderation

Research about the role of anonymity in comment sections of newspaper websites has been prolific. It provides additional insight into anonymity by showing us a picture of how forums that are not interest-specific (like some subreddits) or extremist (like some 4chan message boards) are affected by it.

Several studies seek to evaluate the role of anonymity by assessing discursive civility, which an influential study notes "has been defined as arguing the justice of one's own view while admitting and respecting the justice of others' views."[85] Civility is not, of course, the only value that critics of anonymity online argue that it threatens.[86] Anonymity has also been linked to hate speech, actual threats, and harassment. Nevertheless, research on civility can help shed light on the extent to

[83] PERSILY, *supra* note 6 at 21. See *supra* notes 73-75 and accompanying text.

[84] *See supra* notes 65-69 and accompanying text.

[85] Arthur D. Santana, *Virtuous or Vitriolic: The Effect of Anonymity on Civility in Online Newspaper Reader Comment Boards*, 8 JOURNALISM PRACTICE 18, 21 (2014).

[86] And neither is civility valuable in every given circumstance; it would be unwarranted to expect civility from those who are faced with abuse. Indeed, recent work has criticized the weight given to civility measures as a proxy for deliberative quality. *See* Patrícia Rossini, *Beyond Incivility: Understanding Patterns of Uncivil and Intolerant Discourse in Online Political Talk*, 49 COMM. RES. 399, 400 (2022). The point here is to discuss the incivility-inducing role attributed to anonymity. I make no normative claim about civility, but instead aim to show how those concerned with it would be wrong to assume anonymity is the culprit.

which anonymity drives people to behave without respect for social norms, which include but are not limited to, disapproval of uncivil speech.

What, then, do studies on comment sections tell about civility and anonymity? The evidence is mixed. A highly cited 2014 study compared 11 online newspapers and found that "over 53 percent of the anonymous comments were uncivil, while 28.7 percent of the non-anonymous comments were uncivil."[87] The same researcher more recently examined 30 outlets, with similar results.[88] Yet competing explanations were not discussed, and so differences in the audiences of each website as well as varying content moderation practices could have interfered with the observed effects. Another study compared comments on The Washington Post website, which "afford[ed] users a relatively high level of anonymity," with the newspaper's Facebook page, finding the former had significantly more uncivil discussions than the latter.[89] Yet again other factors cannot be excluded, and it is plausible that content moderation practices in early 2013 available to and deployed by Facebook were considerably more efficient than those The Washington Post website could make use of. Conversely, a study comparing comments posted to newspaper websites and respective Facebook pages in Brazil in 2016 (a period of considerable disruption that saw President Dilma Rousseff's removal from office after her impeachment trial) identified no significant difference in terms of incivility and actually found more intolerance on Facebook.[90]

Knustad and Johansson examined the toxicity of the comments section of The New York Times and The Washington Post and assessed whether anonymous commenters were more toxic than non-anonymous commenters.[91] The outlets were selected for comparison because they are both "east-coast, national, fairly mainstream, left-leaning newspapers," thus reducing "the likelihood of interfering

[87] Santana, *supra* note 85 at 27.

[88] Arthur D. Santana, *Toward Quality Discourse: Measuring the Effect of User Identity in Commenting Forums*, 40 NEWSPAPER RES. J. 467 (2019).

[89] Ian Rowe, *Civility 2.0: A Comparative Analysis of Incivility in Online Political Discussion*, 18 INFO. COMM. SOC'Y 121 (2014).

[90] Rossini, *supra* note 86 at 416 ("[P]latform was not a significant predictor of incivility . . . . Differently than incivility, intolerance is more likely to be expressed on Facebook.").

[91] Magnus Knustad & Christer Johansson, *Anonymity and Inhibition in Newspaper Comments*, 12 INFO. 106 (2021).

variables, such as the affordances of different platforms, with different rules of conduct, moderation and different comment section cultures."[92] They found a "small or tiny" correlation between anonymity and toxic comments, but a much larger difference between the two publications. The Post had considerably more toxic comments than The Times comments section. This led researchers to conclude that "website is a stronger explanation for toxicity than anonymity alone."[93] The authors speculated that these results might be a product of different content moderation strategies since both newspapers "have extensive community rules and guidelines that are linked to in the comment sections . . . that reflect their desire for civil and well-informed comments, and neither allow personal attacks, vulgarity or off-topic comments,"[94] noting that The Times uses machine learning software developed by Jigsaw,[95] part of the Alphabet conglomerate. The researchers hypothesized The Times' system might be "better at catching unwanted comments than the system used by The Washington Post,"[96] which boasted about having its own, proprietary machine learning system.[97]

Another potential factor is that The Times also banks on "NYT Picks," which are selected by the moderators to showcase "high quality comments with exceptional insights that are highlighted in the commenting interface."[98] A study found evidence of "the positive impact of highlighting desirable behaviors via NYT Picks to encourage a higher-quality communication in online comment communities."[99]

[92] *Id.* at 6.

[93] *Id.* at 12.

[94] *Id.*

[95] Bassey Etim, *The Times Sharply Increases Articles Open for Comments, Using Google's Technology*, N.Y. TIMES (Jun. 13, 2017), https://www.nytimes.com/2017/06/13/insider/have-a-comment-leave-a-comment.html.

[96] Knustad & Johansson, *supra* note 91 at 12.

[97] *The Washington Post Leverages Artificial Intelligence in Comment Moderation*, WASH. POST (Jun. 22, 2017), https://www.washingtonpost.com/pr/wp/2017/06/22/the-washington-post-leverages-artificial-intelligence-in-comment-moderation/.

[98] Deokgun Park et al., *Supporting Comment Moderators in Identifying High Quality Online News Comments*, PROC. 2016 CHI CONF. ON HUM. FACTORS COMPUTING SYS. 1114, 1114 (2016).

[99] Yixue Wang & Nicholas Diakopoulos, *Highlighting High-Quality Content as a Moderation Strategy: The Role of New York Times Picks in Comment Quality and Engagement*, 4 ACM TRANSACTIONS ON SOCIAL COMPUTING 1, 3 (2021). "Our findings include the following: (1) Picks are cor-

The Post also highlighted comments, not for their quality, but to call attention to "[u]sers with direct involvement in a particular story."[100] The Times' content moderation strategy of spotlighting quality contributions while taking advantage of design features might be an important factor in the differences found by research on anonymous comments.

This reaffirms the centrality of content moderation practices to understanding how anonymous communities work. Policies, strategies, and enforcement are crucial in governing the digital public sphere, and not just as assessed by, e.g., the volume or prevalence of infringing or abusive content. The point here is not that creative content moderation or more efficient systems can keep the anonymous vandals out. Indeed, we should not underestimate issues with automation in content moderation,[101] particularly with Perspective, the Jigsaw software which was adapted to create The Times' Moderator.[102] The point is instead that content moderation is a component in shaping the identity of those taking part in a particular

related with an improvement in first-time receivers' next approved comment quality, with the quality boost associated with receiving a Pick attenuating after subsequent Picks; (2) receiving a Pick is associated with commenters early in their tenure on the site (i.e., within their first 2 approved comments) returning to the comment section more quickly to make their next comment; and (3) The quality of the visible commentary is positively associated with the quality of subsequent approved commentary. Exposure to Pick badges is also associated with subsequently writing higher-quality approved reply comments, though to a somewhat lesser degree compared to the impact of the quality of parent comments." Wang & Diakopoulos, *id.* at 19.

[100] *The Washington Post Leverages Artificial Intelligence in Comment Moderation*, *supra* note 97. It later announced "featured comments," "picked by Post staff members to highlight thoughtful and diverse contributions to the discussion." *Community Rules*, WASH. POST (Apr. 13, 2020), https://www.washingtonpost.com/discussions/2020/04/13/community-rules/.

[101] *See* Hannah Bloch-Wehba, *Automation in Moderation*, 53 CORNELL INT'L L.J. 42 (2020); James Grimmelmann, *The Virtues of Moderation*, 17 YALE J. L. & TECH. 42, 63–65 (2015).

[102] *See* Matthew J. Salganik & Robin C. Lee, *To Apply Machine Learning Responsibly, We Use It in Moderation*, N.Y. TIMES (Apr. 30, 2020), https://open.nytimes.com/to-apply-machine-learning-responsibly-we-use-it-in-moderation-d001f49e0644 (discussing biases and limitations of the NYT software); Thiago Dias Oliva, Dennys Marcelo Antonialli, & Alessandra Gomes, *Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online*, 25 SEXUALITY & CULTURE 700 (2020) (finding perspective to evaluate white nationalist speech as less toxic than drag queens'). *See also* Aaron Mendon–Plasek, *Mechanized Significance and Machine Learning: Why It Became Thinkable and Preferable to Teach Machines to Judge the World*, *in* THE CULTURAL LIFE OF MACHINE LEARNING 31, 34–36 (Jonathan Roberge & Michael

digital forum, which takes place even when identification is not required.[103] Content moderation can do so by modeling positive behavior, as with the NYT Picks (or flairs in some subreddits, as seen in Part II, Section B), as well as by curbing unwelcome content and preventing users from being provoked into emulating it.

III. False Information, Polarization, and Identity

Identification is often seen as a means for better democratic deliberation. Anonymity is regarded as an abettor of lying.[104] By contrast, as "more information moves the market closer to truth,"[105] identification makes for an improved marketplace of ideas, by equipping listeners with better information to form their judgment.[106] Identification is taken "as a beneficial and purifying process,"[107] through which "[t]he sense of being exposed to public view spurs us to engage in the actions

Castelle eds., 2021) (discussing limitations in the approach that the developers of Perspective API adopted to create a toxicity classifier).

[103] See Tushnet, *supra* note 19 at 108 ("Instead of focusing on names, online discourse would be better served by comment moderation or other forms of curation that can operate to serve similar purposes as norms of behavior in physical public spaces, where we likewise don't usually know legal names but nonetheless generally expect certain constraints to hold."). Note however that Tushnet emphasizes a contrast between pseudonymity and anonymity and sees a role for community building for the former. *See id.* at 84. The argument here is consistent with her points on persistent pseudonyms but expands them to anonymity. See Monteiro, *supra* note 5 at 77–81 for sources and a discussion of situated anonymity and anonymous intimacy in settings of non-persistent pseudonyms.

[104] *See* 514 U.S. 334, 382 (1995) (Scalia, J., dissenting) ("I am sure . . . that a person who is required to put his name to a document is much less likely to lie than one who can lie anonymously."). *See also* Persily, *supra* note 6 at 16 ("Anonymity and pseudonymity (adopting an online persona other than one's own) also facilitate the kind of lying and misrepresentation that undercut a well-informed electorate. In the internet world, anonymous and pseudonymous speakers cannot be held to account for the truth of their electorally relevant statements. Consequently, the speaker bares no cost for repeating lies and promoting false content.").

[105] Kreimer, *supra* note 17 at 74.

[106] Frederick Schauer, *Anonymity and Authority*, 27 J.L. & Pol. 597, 606 (2012) ("The identity of a speaker, and the signals about reliability that may be provided by knowing the speaker's identity, are part and parcel of the content of what a speaker says and of how listeners evaluate it.").

[107] Kreimer, *supra* note 17 at 89 (describing arguments in favor of disclosure).

of the person we would like to be."[108] "[C]ivil and dignified" discourse is also associated with identification,[109] which furthermore upholds civic virtues needed for democratic decision-making.[110]

The previous Part has shown that categorical statements such as those do not appreciate how anonymous settings can shape identities in different ways. Just as anonymity in Reddit and 4chan results in contrasting outcomes, we should not expect that anonymity will always undermine the democratic values with which commentators are concerned. Anonymity is not intrinsically inferior to identification. That is because the effects of anonymity and, as Part I showed, identification vary. This Part goes further than claiming anonymity is not *less than*. I will argue that, in fact, identification can be an agent in the pathologies afflicting social media, particularly dis- and misinformation.

To see how, we need to understand the real-world interplay of identity in its articulation with community and norms. Commentators have assumed that anonymity "facilitate[s] the kind of lying and misrepresentation that undercut a well-

[108] *Id.* at 92.

[109] *See* 514 U.S. 334, 382 (1995) (Scalia, J., dissenting): "[T]he usefulness of a signing requirement lies not only in promoting observance of the law against campaign falsehoods (though that alone is enough to sustain it). It lies also in promoting a civil and dignified level of campaign debate—which the State has no power to command, but ample power to encourage by such undemanding measures as a signature requirement." A Pew Research Center report on the results of "a large-scale canvassing of technology experts, scholars, corporate practitioners, and government leaders" found that "many . . . attributed [anonymity] to the enabling bad behavior and facilitating 'uncivil discourse' in shared online spaces." LEE RAINIE, JANNA ANDERSON, & JONATHAN ALBRIGHT, THE FUTURE OF FREE SPEECH, TROLLS, ANONYMITY AND FAKE NEWS ONLINE 3–4 (2017), https://www.pewresearch.org/internet/2017/03/29/the-future-of-free-speech-trolls-anonymity-and-fake-news-online/. Views on anonymity included: "People are snarky and awful online in large part because they can be anonymous, or because they don't have to speak to other people face-to-face." "Anonymity (or at least the illusion of it) feeds a negative culture of vitriol and abuse that stifles free speech online." RAINIE, ANDERSON, & ALBRIGHT, *id.* at 36. "Increased anonymity coupled with an increase in less-than-informed input, with no responsibility by the actors, has tended and will continue to create less open and honest conversations and more one-sided and negative activities." Rainie, Anderson, & Albright, *id.* at 8. *See also* PERSILY, *supra* note 6 at 16.

[110] *See* Kreimer, *supra* note 17 at 101–02 ("First, publicity assures the quality of debate by acting as a check on the qualities of the debaters. . . . Second, publicity improves the character and judgement of the citizenry. Open participation in public life exercises and develops the virtue of courage."). But *see* Kreimer, *id.* at 107 (arguing that such virtues associated with identification require "the concrete analysis of the situations in which claims of anonymity are exerted.").

informed electorate" because "the speaker bares no cost for repeating lies and promoting false content."[111] But research into political polarization paints a different picture. It tells us that, in affectively polarized settings,[112] identified speakers can reap rewards from lies and inaccuracies, instead of being punished by their listeners.

In an affectively polarized landscape where hyperpartisans dominate social media,[113] accuracy and truth do not take the disciplining force that the commentary about identification assumes.[114] In fact, users have been shown to sever the decision to share from their judgment on the truthfulness or falsity of the news.[115] This indicates that it is not only anonymity but also identification that has been insufficiently conceptualized. The next sections will bring together findings from different strands of scholarly literature to explore the role that identification plays in the sharing of mis- and disinformation.

[111] PERSILY, *supra* note 6 at 16.

[112] Iyengar et al. describe affective polarization in terms of animosity between political parties. Shanto Iyengar et al., *The Origins and Consequences of Affective Polarization in the United States*, 22 ANN. R. POL. SCI. 1, 130 (2018) ("Democrats and Republicans both say that the other party's members are hypocritical, selfish, and closed-minded, and they are unwilling to socialize across party lines, or even to partner with opponents in a variety of other activities. This phenomenon of animosity between the parties is known as affective polarization."). *See also infra* notes 123-125.

[113] *See*, e.g., CHRIS BAIL, BREAKING THE SOCIAL MEDIA PRISM 76 (2021) (citation omitted): "A 2019 report from Pew showed that a small group of people is responsible for most political content on Twitter. Specifically, this report found that 'prolific political tweeters make up just 6% of all Twitter users but generate 20% of all tweets and 73% of the tweets mentioning national politics.' What is more, extremists represented nearly half of all prolific tweeters. Though people with extreme views constitute about 6 percent of the U.S. population, the Pew report found that '55% of prolific political tweeters identity as very liberal or very conservative.'"

[114] *See* Mathias Osmundsen et al., *Partisan Polarization Is the Primary Psychological Motivation Behind Political Fake News Sharing on Twitter*, 115 AM. POL. SCI. REV. 999, 1012 (2020): "From a partisan-motivated perspective, fake news is not categorically different from other sources of political information. . . . [P]artisans' decisions to share both fake and real news sources depend on how politically useful they are in derogating the out-party."

[115] Gordon Pennycook & David G. Rand, *The Psychology of Fake News*, 25 TRENDS COGNITIVE SCI. 388, 6 (2021) (". . . participants who were asked about the accuracy of a set of headlines rated true headlines as much more accurate than false headlines; but, when asked whether they would share the headlines, veracity had little impact on sharing intentions . . . ."); Gordon Pennycook et al., *Shifting Attention to Accuracy Can Reduce Misinformation Online*, 592 NATURE 590 (2020).

A. Identity and Affective Polarization

An important concern about the current state of the online landscape is political polarization, which has been described as "the greatest threat to American democracy"[116] and one of the "four horsemen of constitutional rot."[117] Social media has been blamed for reinforcing preexisting beliefs through repeated exposure to homogeneous viewpoints, which in turn further cements beliefs and insulates them from being challenged. It thus contributes to increasingly polarized politics, with each side of the divide living in its own "echo chamber," according to a popular account of the issue.[118]

In fact, the "echo chambers" theory of social media as a driver of polarization is quite controversial. Researchers have found little empirical support for the thesis, or have concluded that the claim is overstated.[119] A study has found that modest monetary incentives may considerably dissipate reported incorrect partisan beliefs about facts.[120] And even when it comes to opinions, the echo chambers account might fail to consider how partisan attitudes toward policy positions are formed. For instance, in a study, participants voiced support for a policy aligned with their perception of party ideology but expressed the contrary view when told party stance was in favor of the policy.[121] Furthermore, the echo chamber account of political

[116] Erwin Chemerinsky, *False Speech and the First Amendment*, 71 OKLA. L. REV. 1, 14 (2017).

[117] *See* JACK BALKIN, THE CYCLES OF CONSTITUTIONAL TIME 49 (2020): "There are four basic causes of constitutional rot—I call them the Four Horsemen of Constitutional Rot. The first is political polarization." (Citation omitted.)

[118] CASS SUNSTEIN, #REPUBLIC (3 ed. 2018).

[119] PABLO BARBERÁ, *Social Media, Echo Chambers, and Political Polarization*, *in* SOCIAL MEDIA AND DEMOCRACY: THE STATE OF THE FIELD, PROSPECTS FOR REFORM 34, 35–41 (Nathaniel Persily & Joshua A. Tucker eds., 2019); ANDREW GUESS ET AL., AVOIDING THE ECHO CHAMBER ABOUT ECHO CHAMBERS: WHY SELECTIVE EXPOSURE TO LIKE-MINDED POLITICAL NEWS IS LESS PREVALENT THAN YOU THINK (2018).

[120] John G Bullock et al., *Partisan Bias in Factual Beliefs About Politics*, Q.J. POL. SCI. 519 (2015).

[121] *See* Geoffrey L. Cohen, *Party over Policy: The Dominating Impact of Group Influence on Political Beliefs*, 85 J. PERSONALITY & SOC. PSYCHOL. 808, 819 (2003) (summarizing the findings: "If information about the position of their party was absent, liberal and conservative undergraduates based their attitude on the objective content of the policy and its merit in light of long-held ideological beliefs. If information about the position of their party was available, however, participants assumed that position as their own regardless of the content of the policy.").

polarization may dramatically underestimate the role of legacy media actors in driving the phenomenon.[122]

Instead of issue-based, or ideological, polarization, many scholars are increasingly more interested in the escalation of *affective polarization*, described as a "phenomenon of animosity between the parties."[123] Affective polarization refers to the process whereby identities get sorted along a cultural divide that predicts where people buy food, what clothes they wear, and what shows they watch on TV.[124] In short, rather than policy positions (e.g., support for a proposed gun control measure), this kind of polarization manifests itself in more encompassing terms. The divide is not only along partisan, but also racial, religious, cultural, and geographic lines,[125] all of which are increasingly conflated.

[122] YOCHAI BENKLER, ROBERT FARRIS, & HAL ROBERTS, NETWORK PROPAGANDA 386 (2018) ("There is no echo chamber or filter-bubble effect that will inexorably take a society with a well-functioning public sphere and turn it into a shambles simply because the internet comes to town. The American online public sphere is a shambles because it was grafted onto a television and radio public sphere that was already deeply broken. Even here, those parts of the American public sphere that were not already in the grip of a propaganda feedback loop and under the influence of hyperpartisan media dedicated to a propagandist project did not develop such a structure as a result of the internet's development.").

[123] *See* Iyengar et al., *supra* note 112 at 130: "Democrats and Republicans both say that the other party's members are hypocritical, selfish, and closed-minded, and they are unwilling to socialize across party lines, or even to partner with opponents in a variety of other activities. This phenomenon of animosity between the parties is known as affective polarization." *See also* Shanto Iyengar, Gaurav Sood, & Yphtach Lelkes, *Affect, Not Ideology*, 76 PUB. OP. Q. 405 (2011); Lilliana Mason, *The Rise of Uncivil Agreement: Issue Versus Behavioral Polarization in the American Electorate*, 57 AM. BEHAV. SCIENTIST 140 (2013).

[124] *See* Lilliana Mason, *Losing Common Ground: Social Sorting and Polarization*, 16 THE FORUM 47, 49 (2018): ". . . American partisans are speaking different languages, misunderstanding one another, and distrusting their fellow Americans on a basic level. Where Democrats and Republicans could at one time discuss last night's television shows around the water cooler, today they are not only watching different shows, but they are also drinking different beverages." *See also* LILLIANA MASON, UNCIVIL AGREEMENT: HOW POLITICS BECAME OUR IDENTITY (2018).

[125] *See* Eli J. Finkel et al., *Political Sectarianism in America*, 370 SCIENCE 533, 535 (2020): "Compared to a few decades ago, Americans today are much more opposed to dating or marrying an opposing partisan; they are also wary of living near or working for one. They tend to discriminate, as when paying an opposing partisan less than a copartisan for identical job performance or recommending that an opposing partisan be denied a scholarship despite being the more qualified applicant."

Affective polarization helps explain seemingly paradoxical results from a research intervention that was designed to decrease "echo chamber" insulation (and hence partisan distance). In that study, partisans were paid to follow a Twitter bot account that exposed them to opposing political ideologies.[126] If greater political polarization is understood as the result of social media reinforcing views and information, and not exposing partisans to different thinking, we would expect that participants who saw more cross-partisan content would hold less polarized attitudes. The study instead found that participants subsequently exhibited *more* partisan attitudes.[127] The key to unraveling this paradox is in understanding how identities are shaped on social media in a context of affective polarization.

The echo chamber account sees polarization as a consequence of insulation created by social media. Scholarship highlighting affective polarization instead frames it as being "driven by conflict rather than isolation."[128] Exposure to cross-party content such as offered by the study thus does not break echo chambers, argues the lead author of the study in subsequent work,[129] because it does not breed reflection and deliberation. Rather, it "sharpen[s] the contrasts between 'us' and 'them,'"[130] magnifying affective polarization.

Chris Bail uses the metaphor of a prism to explain how social media plays an important role in shaping political identities in the reflection of a distorted image of society.[131] In a setting where affective polarization festers, extremists get validation and social support from detracting the out-party, as well as from disciplining

[126] The authors "created a liberal Twitter bot and a conservative Twitter bot for each of our experiments. These bots retweeted messages randomly sampled from a list of 4,176 political Twitter accounts (e.g., elected officials, opinion leaders, media organizations, and nonprofit groups)." Christopher A. Bail et al., *Exposure to Opposing Views on Social Media Can Increase Political Polarization*, 115 PROC. NAT'L ACAD. SCI. 9216, 9217 (2018).

[127] The effect was not uniform across political lines: Democrat participants showed slight effects which were not statistically significant, while "Republicans, by contrast, exhibited substantially more conservative views." *Id.* Importantly, "[e]xposing people to views of the other side did not make participants more moderate." BAIL, *supra* note 113 at 20.

[128] Petter Törnberg, *How Digital Media Drive Affective Polarization Through Partisan Sorting*, 119 PROC. NAT'L ACAD. SCI. e2207159119, 10 (2022).

[129] *See* BAIL, *supra* note 113.

[130] *Id.* at 39.

[131] *See id.* at 10 (introducing the social media prism).

in-party members who stray from in-party views. This sort of behavior is then normalized by moderates (who are given the impression that their views are less prevalent than they in fact are)[132] and extremists (who are entrenched further not just in their views but their tactics).[133]

Affective polarization offers identity, more than policy positions or information, as crucial to understanding political polarization on the internet. Platforms magnify feedback processes of the presentation of the self; they "enable us to make social comparisons with unprecedented scale and speed."[134] That is, we can clearly see what sort of content gets positive engagement from other users, and what sort of content brings about the embarrassment of being ignored or the stress of being contested. Given that social media is now so ingrained in everyday life, straying from partisan expectations is very costly, socially and emotionally.[135] This, Bail contends, creates a prism distorting our sense of the environment, inducing us to see the partisan out-group as more extreme than it actually is, through the rewarding of radical partisan behavior and silencing of moderate behavior. All of that culminates in "status seeking on social media creat[ing] a vicious cycle of political extremism."[136]

B. Performing Lies and Misinformation: Identification as a Driver

So social media is a cog in a machine that rewards greater affective polarization. Platforms "do not isolate us from opposing ideas; *au contraire*, they throw us into a national political war."[137] The prevalence of dis- and misinformation online must

[132] *See id.* at 82–83: ". . . the social media prism makes the other side appear monolithic, unflinching, and unreasonable. While extremists captivate our attention, moderates can seem all but invisible."

[133] *See id.* at 66–67 (describing "extremism through the prism").

[134] *Id.* at 51.

[135] *See id.* at 77: "Posting online about politics simply carries more risk than it's worth. Such moderates [*as opposed to with extremists*] are keenly aware that what happens online can have important consequences off-line."

[136] *See id.* at 53. "Moderates disengage from politics on social media for several different reasons. Some do so after they are attacked by extremists. Others are so appalled by the breakdown in civility that they see little point to wading into the fray. Still others disengage because they worry that posting about politics might sacrifice the hard-fought status they've achieved in their off-line lives." *Id.* at 83.

[137] Törnberg, *supra* note 128 at 10.

be understood against that background. Once we appreciate this, the connection between identification (particularly the kind established by real-name policies) and misinformation is made clear.

There is increasing evidence that content employing "moral-emotional language" does significantly better on social media. Moral psychologists use the term "moral-emotional language" to refer to language which both expresses a moral judgment about what is right and wrong and an emotional state (such as "hate" or "contempt").[138] Moral-emotional content shows a propensity to go viral online, in a process researchers have described as "moral contagion" given how "it mimics the spread of disease."[139] One study of over 500,000 tweets found a 20 percent increase in sharing for each word marked by that kind of language.[140]

Disinformation campaigns have leveraged that viral propensity of moral-emotional content.[141] A study that looked at news articles shared on Twitter concluded false news (established as such through concurring assessments by fact-checking organizations) evoked disgust, more so than real news.[142]

[138] See William J Brady et al., *Emotion Shapes the Diffusion of Moralized Content in Social Networks*, 114 PROC. NAT'L ACAD. SCI. 7313, 7313 (2017) (describing moral-emotional language). Note that "moral expression" is employed in the broadest possible terms, with reference to "what is perceived as 'right' and 'wrong.'" Gun control is given an example of "moralized content," and contrasted with "a social-media message about cute kittens." William J. Brady, M. J. Crockett, & Jay J. Van Bavel, *The MAD Model of Moral Contagion: The Role of Motivation, Attention, and Design in the Spread of Moralized Content Online*, 15 PERSPECTIVES ON PSYCHOL. SCI. 978, 978–79 (2020).

[139] Brady, Crockett, & Van Bavel, *supra* note 138 at 7313.

[140] *See id.* at 7316: "Using a large sample of tweets concerning three polarizing issues (n = 563,312), the presence of moral-emotional words in messages increased their transmission by approximately 20% per word."

[141] "[M]oral and emotional appeals that capture attention can be exploited by disinformation profiteers, as in the case of fake news spread around the 2016 U.S. election[.]" Brady, Crockett, & Van Bavel, *supra* note 138 at 20.

[142] "Whereas false stories inspired fear, disgust, and surprise in replies, true stories inspired anticipation, sadness, joy, and trust." Soroush Vosoughi, Deb Roy, & Sinan Aral, *The Spread of True and False News Online*, 359 SCIENCE 1146, 1146 (2018).

Research focused on moral outrage (a subcategory of moral-emotional language)[143] on social media has explored the mechanisms behind the virality of that sort of content. I want to foreground how identification is a part of those mechanisms.

One important component in expressing moral outrage is the reputational gains that can be reaped by signaling to the in-group that we care about serious moral violations.[144] This logic is valid both offline and online, but whereas expressing outrage at, for instance, how badly a fellow commuter was treated might typically get us credit with others waiting in the subway station, for instance, "doing so online instantly advertises your character to your entire social network and beyond," as M. J. Crockett puts it.[145] This line of research emphasizes how social media "is a context in which our political group identities are hypersalient,"[146] which amplifies motivations to engage in moral-emotional expression to uphold in-group versus perceived out-group threats[147] and to accrue in-group reputational gains.[148] Status seeking (much like Bail described) is an important component of the spread of dis- and misinformation online.

When users are anonymous in online settings, they are less likely to express outrage,[149] as an important part of their underlying personal motivation is removed:

[143] Moral outrage is defined as "an intense negative emotion combining anger and disgust triggered by a perception that someone violated a moral norm." Jordan Carpenter et al., *Political Polarization and Moral Outrage on Social Media*, 52 CONN. L. REV. 1107, 90 (2021).

[144] M. J. Crockett, *Moral Outrage in the Digital Age*, 1 NATURE HUMAN BEHAV. 769, 770 (2017) (noting that "expressing moral outrage benefits individuals by signalling their moral quality to others").

[145] *Id.*

[146] Brady, Crockett, & Van Bavel, *supra* note 138 at 989.

[147] *See id.* at 985: "When out-group members pose threats to the moral values of the in-group, out-group derogation is a common in-group response to uphold a positive in-group image . . . . In other words, condemning an out-group's behavior makes one's in-group appear better by comparison."

[148] *Id.* at 987: "[E]xpressing moral emotions that derogate the out-group or bolster the in-group can enhance one's reputation and increase group belonging."

[149] *See id.*: "For example, in online settings people are more likely to express outrage toward policies they oppose when their identity is *not* anonymous, suggesting that the opportunity to signal to others should be associated with a greater likelihood of expressing outrage online" (emphasis in the original, citation omitted).

"the need to maintain an image as a good group member in the eyes of other group members."[150] This is supported by related research on "online aggression" in a German petitions website, which found that non-anonymous users comments were *more* aggressive when engaging in firestorms[151] against public officials and policies, given that aggressiveness would not be something they would want to conceal—on the contrary, they would want to be seen as standing up for their values.[152]

My argument is that real names on social media significantly raise the stakes of the rewards for expressing moral outrage. Granted, pseudonymous users can benefit from reputational gains within that particular digital context. The reputational gains for *identified* users, however, can render them material benefits, including offline. The favorable recognition they can achieve might be translated, for instance, in media appearances in prestigious legacy publications or in professional opportunities. The fact that they can extract those tangible gains is a function of the affordances for identification in a given platform. A platform that disfavors identification, where users are not identified across different posts, like YikYak or Whisper, impedes users who might be willing to claim to be the authors of a viral piece of content; they will have problems establishing themselves as the genuine posters—anyone would be able to fabricate a screenshot and try to get credit.

The entanglement between identification and moral outrage goes further than the rewards those expressing it online can garner. Moral outrage directed at a member of the out-group upholds in-group norms and thus also affirms group identity,

[150] *Id.* at 995. "In other words, expressing moral emotions that derogate the out-group or bolster the in-group can enhance one's reputation and increase group belonging." *Id.*

[151] *See* Katja Rost, Lea Stahel, & Bruno S. Frey, *Digital Social Norm Enforcement: Online Firestorms in Social Media*, 11 PLOS ONE 1, 2 (2016) (citations omitted): "In online firestorms, large amounts of critique, insulting comments, and swearwords against a person, organization, or group may be formed by, and propagated via, thousands or millions of people within hours. Social media enable these unleashed phenomena. They allow attacking everywhere at anytime with the potential for an unlimited audience."

[152] *See id.* at 17: "[O]nline anonymity does not promote online aggression in the context of online firestorms. There are no reasons for anonymity if people want to stand up for higher-order moral principles and if anonymity decreases the effectiveness of sanctions for norm enforcement." *See also* Lea Stahel & Katja Rost, *Angels and Devils of Digital Social Norm Enforcement*, PROC. 8TH INT'L CONF. ON SOC. MEDIA & SOC'Y 17 1, 6 (2017): "[Users enforcing norms in online firestorms] comment more aggressively . . . if they comment non-anonymously . . . ."

to the detriment of the out-group.[153] In an affectively polarized setting, moral outrage is an assault on the opposing party and its political capital. Identification again is crucial here. Real names in a polarized setting will enable and invite users to try to establish whether the target of their outrage is a member of the opposing party. Digital encounters in identified settings then provide opportunities, particularly for hyperpartisans, to raid the opposing party at every flank where moral outrage can be expressed. Even if in a given platform users find insufficient cues about the potential targets for moral outrage expression (such as how they identify through their bios, their profile picture, likes, or follows), other information on the web can be found to try and infer party affiliation.

To be clear, this kind of antagonistic behavior is not exclusive to real-name settings; it can also take place with pseudonyms whenever there are sufficient cues for users to make inferences about others. Yet this mechanism is contingent on the norms in an anonymous setting: it depends, that is, on whether or not users do bear their party affiliations or give them away inadvertently. In real-name social media, platform design makes this inescapable. As noted earlier, these in-group-oriented motivations extend to the sharing of fake news, regardless of whether the user "ha[s] a firm belief in" it, as research has found.[154] Indeed, one line of study highlights that whether or not people believe false information stands separately from whether they condone it—"they recognize it as false, but give it a moral pass."[155]

And while it might be objected that moral outrage did not start with the internet, M.J. Crockett points to several factors explaining why outrage is amplified by

[153] *See* Brady, Crockett, & Van Bavel, *supra* note 138 at 985: "When out-group members pose threats to the moral values of the in-group, out-group derogation is a common in-group response to uphold a positive in-group image . . . . In other words, condemning an out-group's behavior makes one's in-group appear better by comparison."

[154] *See* Pennycook et al., *supra* note 115 at 594 ("[W]e found a dissociation between accuracy judgments and sharing intentions that suggests that people may share news that they do not necessarily have a firm belief in."); *see also* Pennycook & Rand, *supra* note 115 at 6 ("[P]articipants who were asked about the accuracy of a set of headlines rated true headlines as much more accurate than false headlines; but, when asked whether they would share the headlines, veracity had little impact on sharing intentions . . . .").

[155] Daniel A. Effron & Beth Anne Helgason, *The Moral Psychology of Misinformation: Why We Excuse Dishonesty in a Post-Truth World*, 47 CURRENT OP. PSYCHOL. 101375, 1 (2022).

social media.[156] It multiplies opportunities: There is evidence that in-person observation of violation of moral norms is uncommon.[157] In platforms driven by user engagement, moral outrage is more likely to go viral. And while expressing moral outrage in person is costly (because many will shy away from confrontation or be intimidated by the risk of retaliation from the target of outrage, including with violence), online the costs are lower,[158] and the corresponding positive feedback can be much more immediate.[159] Again, such positive feedback for the individual translates into gains to their reputation, accrued in terms of virtue signaling to the in-group, which is a function of their identification.[160] Once more, online identification with real names can yield different results compared to offline identification, as discussed in Part I.

While the discussion so far has emphasized deleterious effects of moral outrage, the literature has emphasized that such emotional phenomena should not be viewed as intrinsically positive or negative, citing for instance the role it has in propelling collective action around activism around social inequality and injustice and

[156] Crockett, *supra* note 144.

[157] *See* Wilhelm Hofmann et al., *Morality in Everyday Life*, 345 SCIENCE 1340, 1341 (2014) (describing the results of a study in which participants reported their daily experiences; less than 5% of those reports were for witnessing or being the target of "immoral acts").

[158] See Crockett, *supra* note 144 at 770 ("Expressing moral outrage can be costly. Offline, moralistic punishment carries a risk of retaliation. But online social networks limit this risk.").

[159] See *id.* ("Of course, online social networks massively amplify the reputational benefits of outrage expression. While offline punishment signals your virtue only to whoever might be watching, doing so online instantly advertises your character to your entire social network and beyond. A single tweet with an initial audience of just a few hundred can quickly reach millions through viral sharing—and outrage fuels virality.")

[160] This should not be overstated. The claim here is not that signaling dynamics play no part in anonymous settings. In experiments of one-shot interactions, after observing selfish behavior, some participants (who performed worse on cognitive tests) still reported (to experimenters) anger, desire for punishment and moral reprobation. Researchers suggest this is because of the role of reputational heuristics, that is, what the reputational stakes individuals think are typically at stake (even if not present in a given setting). This was supported by the fact that participants who performed better in cognitive tests were less likely to act on moral outrage when punishing selfishness was costly to them. *See* Jillian J. Jordan & David G. Rand, *Signaling When No One Is Watching: A Reputation Heuristics Account of Outrage and Punishment in One-Shot Anonymous Interactions*, 118 J. PERSONALITY & SOC. PSYCHOL. 57 (2019). Note, however, that this study did not examine settings dominated by political polarization.

fundraising campaigns, for instance.[161] The point here is not to pass judgment on moral outrage but to note its part in the mechanisms that underpin the sharing of misinformation online and highlight how identification magnifies those mechanisms.

Commentators see identification as beneficial because they believe users will behave better by refraining from toxic speech out of fear of how their actions online will impact their standing in their social circles.[162] What is generally not accounted for in that narrative is how real names on social media also impel users to *perform* their context-collapsed identities under a condition of affective polarization. The audience (composed of their friends, family, coworkers, and so on) is watching and will pass judgment on deviations from group loyalties. In real-name platforms, experimentation, self-questioning, and crossing the aisle to try to understand the other side come at a price. Posts and comments supporting the in-group are rewarded; in-group opposing content will often lead to disciplining. Risks flowing from context-collapsed identities in social media have been described in terms of what users will or will not post.[163] What we are considering here is how norm enforcement will effectively shape not only what users themselves post but how they consume content by other users. In other words, the content of the posts users share and how they read posts by others are both in part a function of how identities are presented in a platform. This can create a vicious cycle. Conversely, these drivers can be prevented in certain anonymous settings. This is exactly what some researchers have been exploring and is the topic of Part III, Section C.

[161] Victoria L. Spring, C. Daryl Cameron, & Mina Cikara, *The Upside of Outrage*, 22 TRENDS COGNITIVE SCI. 1067 (2018); Victoria L. Spring, C. Daryl Cameron, & Mina Cikara, *Asking Different Questions About Outrage: A Reply to Brady and Crockett*, 23 TRENDS COGNITIVE SCI. 79 (2018).

[162] *See* PERSILY, *supra* note 6 at 16 ("The norms of civility, the fears of retaliation and estrangement, as well as basic psychological dynamics of reciprocity that might deter some types of speech when the speaker and audience know each other – all are retarded when the speech is separated from the speaker, as it is online.").

[163] Marwick and boyd, *supra* note 29 at 122 (describing how context collapse "creates a lowest-common denominator effect," where users will avoid topics they think may alienate their followers).

C. Anonymity as a Depolarizing, Discourse-Enabling Device

With polarization breaking records[164] and social media engulfed in a vicious cycle elicited by status seeking based on constant feedback from like-minded individuals, it might sound like an inane notion to participate in online communities to solicit views contradicting our beliefs on topics such as immigration, gender identity, and the disbandment of one of the main political parties in the U.S. Still, those are examples of conversations at r/ChangeMyView,[165] a subreddit created in 2013 to serve as a venue where users deliberately invite challenges to their opinions.[166]

The community operates within Reddit, which, as discussed above, requires no more than a username and password for account creation and employs a policy that allows and even encourages temporary or "throwaway" accounts.[167] It illustrates how anonymity, combined with platform design and content moderation strategies, can mold identity to support digital spaces in overcoming afflictions plaguing much of social media.

[164] In 1960, about 5% of survey respondents stated they would be displeased if their child married someone from the opposing party. By 2010, roughly one-third of Republicans and half of Democrats expressed they were somewhat upset or very upset by the prospect. *See* Iyengar, Sood, & Lelkes, *supra* note 123 at 416–18. Ahead of the 2022 U.S. midterm elections, the Pew Research Center found that 62% of Republicans expressed very unfavorable views of Democrats, with Democrats reporting at 54%. That was up from 21% and 17% respectively in 1994. It also found all-time highs for respondents describing members of the other party as immoral, dishonest, unintelligent, lazy and closed-minded. *See* PEW RESEARCH CENTER, AS PARTISAN HOSTILITY GROWS, SIGNS OF FRUSTRATION WITH THE TWO-PARTY SYSTEM (2022), https://www.pewresearch.org/politics/wp-content/uploads/sites/4/2022/08/PP_2022.09.08_partisan-hostility_REPORT.pdf (last visited Feb. 27, 2023).

[165] *See* SHAGUN JHAVER, PRANIL VORA, & AMY BRUCKMAN, DESIGNING FOR CIVIL CONVERSATIONS: LESSONS LEARNED FROM CHANGEMYVIEW 4 (2017) (providing examples of posts at r/ChangeMyView).

[166] *See Wiki, r/ChangeMyView*, REDDIT (2018), https://www.reddit.com/r/changemyview/wiki/ (last visited Feb. 27, 2023) ("What is /r/changemyview? . . . CMV is the perfect place to post an opinion you're open to changing."). *See also* Chenhao Tan et al., *Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-Faith Online Discussions*, PROC. 25TH INT'L WORLD WIDE WEB CONF. 613, 613 (2016) (brief description of the subreddit by one of the first works on it); JHAVER, VORA, & BRUCKMAN, *supra* note 165 at 3–4 (describing the community and quoting the subreddit's creator as stating that the "his goal behind creating CMV was not to facilitate debates but to motivate conversations that help users understand different perspectives.").

[167] *See* Leavitt, *supra* note 57 at 320 (describing throwaway accounts).

First, rather than all-encompassing policies that must apply to a wide range of contexts,[168] at r/ChangeMyView the rules meticulously govern not just what can and cannot be posted but also how.[169] They cover the text,[170] the attitude,[171] and the manner and effort of participation[172] expected from users who submit issues to the community. The rules are accompanied by "indicators of violations," which give more insight into how the rules are interpreted and applied.[173] There are also rules for commenters, establishing, for example, that top-level comments (i.e., direct responses to the OP) "must challenge or question at least one aspect of the submitted view," whereas comments within a thread may express agreement with the OP.

Second, the subreddit also leverages platform design to promote the community's goals. The rules also set out criteria for when to award and when not to award deltas, which any user is able to do. Deltas are "a token of appreciation towards a user who helped tweak or reshape your opinion."[174] Deltas are displayed as community badges within r/ChangeMyView. Like mainstream social media, then, the subreddit makes use of gamification strategies;[175] unlike them, however, it does not

[168] *See* Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598, 1642 (2018) (quoting from an interview with former Facebook employees who developed the first set of policies stating "There are no 'places' in Facebook—there are just people with different nationalities, all interacting in many shared forums.").

[169] *Rules, r/ChangeMyView*, REDDIT (2023), https://www.reddit.com/r/changemyview/wiki/rules/ (last visited Feb. 27, 2023).

[170] Rule A for submissions states that original posters ("OPs") must "explain the reasoning behind [their] view, not just what that view is," and that this elaboration requires at least 500 characters. *Id.* Under Rule C, "Submission titles must adequately sum up your view and include 'CMV:' [for 'change my view'] at the beginning." Rule D specifies that "[p]osts cannot express a neutral stance, suggest harm against a specific person, be self-promotional, or discuss this subreddit." *Id.*

[171] Rule B insists that OPs must "personally hold the view and demonstrate" that they are "open to changing it." *Id.*

[172] Rule E requires the OP to be "willing to have a conversation" and "available to do so within 3 hours after posting." *Id.*

[173] For instance, a Rule E violation is assessed, resulting in removal of the post, if the OP does respond within the three-hour period yet only engages the conversation with "[a] small number of one line responses that don't address the arguments that people are making." *Id.*

[174] *Id.*

[175] *See* JHAVER, VORA, & BRUCKMAN, *supra* note 165 at 4 ("The community gamifies the process of changing the view of post submitters by implementing an award mechanism called the delta system").

optimize for user engagement, and instead leverages platform affordances to "celebrat[e] view changes, [which] is at the core of Change My View."[176]

Third, policy enforcement. Moderators are active and adopt a range of approaches to steer the subreddit.[177] An extensive set of "Moderation standards and practices" addresses "procedures for removing posts/comments, how bans are decided and implemented, how the six (6) month statute of limitations is applied for offenses, and how our appeal process works."[178] Policy enforcement is therefore also tailored to support the community's goals, including providing explanations for post removals,[179] adapting automation tools,[180] and employing and modifying design features,[181] such as flags and the delta system.

The extent to which r/ChangeMyView actually vindicates its name is debated. Researchers conducted interviews with 15 participants, reporting users "typically did not change their view completely,"[182] even though they saw the community as useful. More importantly for affective polarization concerns, they found participants thought "posting on CMV helped them develop empathy towards users they earlier disagreed with."[183]

Another example of anonymity being put to use to achieve what real names could not is DiscussIt, a "mobile chat app [developed] to conduct a field experiment

[176] *Rules, r/ChangeMyView*, *supra* note 169.

[177] *See* JHAVER, VORA, & BRUCKMAN, *supra* note 165 at 6 (reporting that many participants interviewed by the authors "felt that a strict enforcement of [the] rules has been critical in maintaining the civil nature of conversations").

[178] *Moderation Standards and Practices, r/ChangeMyView*, REDDIT (2023), https://www.reddit.com/r/changemyview/wiki/modstandards/ (last visited Feb. 27, 2023).

[179] *See* Kumar Bhargav Srinivasan et al., *Content Removal as a Moderation Strategy: Compliance and Other Outcomes in the ChangeMyView Community*, 3 PROC. ACM HUM.–COMPUT. INTERACTION 1, 4 (2019) (providing an example of the notice explaining reasons for removal of a post).

[180] *See* Chandrasekharan et al., *supra* note 61 at 22 (discussing the use of automated moderation tools by subreddits).

[181] *See* Jisu Kim et al., *Promoting Online Civility Through Platform Architecture*, 1 J. ONLINE TRUST & SAFETY, 15 (2022) (study on Nextdoor, a location-based social media platform, noting that "more civil interactions among users can be encouraged by altering the design and architecture of the online environment within which the interaction occurs").

[182] JHAVER, VORA, & BRUCKMAN, *supra* note 165.

[183] *Id.*

testing the impact of anonymous cross-party conversations on controversial topics."[184] After recruiting 1,200 Democrats and Republicans, DiscussIt matched each of them with a participant from the opposing political party. Participants were issued an androgynous-sounding pseudonym to join a chat with question prompts asking for their views on either immigration or gun control and received notifications if they became non-responsive.[185] Comparing surveys of participants that responded before and after the experiment, one of the authors says he is "cautiously optimistic about the power of anonymity," as "many people expressed fewer negative attitudes to the other parties or subscribed less strongly to the stereotypes about them," and "many others expressed more moderate views about political issues they discussed or social policies designed to address them."[186] The study reported findings of changes in sentiment toward opposing party members as well as views on the issues discussed.[187]

Experiences such as DiscussIt and r/ChangeMyView show us at least two ways that anonymity can be instrumental in creating a more vibrant digital public sphere. One is by attenuating affective polarization, as noted. This is in line with research suggesting that positive contact with the out-group can reduce affective polarization.[188] There is evidence that partisans have exaggerated perceptions of members

[184] Aidan Combs et al., *Anonymous Cross-Party Conversations Can Decrease Political Polarization: A Field Experiment on a Mobile Chat Platform*, SOCARXIV 3 (2022), https://osf.io/preprints/socarxiv/cwgu5/.

[185] *See id.* at 4–7 (discussing research design).

[186] BAIL, *supra* note 113 at 125.

[187] See Combs et al., *supra* note 184 at 9 (discussing findings).

[188] *See* Rachel Hartman et al., *Interventions to Reduce Partisan Animosity*, 6 NATURE HUMAN BEHAV. 1194, 1197–98 (2022) (citations omitted) ("A rich body of literature in social psychology details the positive effects of contact on intergroup relations across barriers related to race, ethnicity, religion and sexual orientation."). *See also* James N. Druckman et al., *(Mis)estimating Affective Polarization*, 84 J. POL. 1106 (2022).

of the opposing party,[189] so that engaging in conversation with a living average Republican or Democrat can disabuse stereotypes and reduce negative attitudes.[190] Anonymous social media can create opportunities for cross-party interaction that do not take place in the battlegrounds of a "national political war,"[191] and where reputation is not gained by scoring points against the opposing party with any available means. We have seen how affective polarization is connected with misinformation, so alleviating one could help with the other.

A further way anonymity can play a part in enacting more truth-based discourse is by enabling conversations that are grounded in facts and guided by what is generally expected of public deliberation—even if it does not move the needle on polarization. Anonymity can lower the stakes of engaging in what could otherwise be seen as heretical partisan equivocation that would be faced with disciplining. It can as such facilitate hard conversations for which at least some users have a hunger.[192]

These examples might be too hopeful if thought of as immediate prototypes for replacing Facebook or Twitter,[193] yet they are still valuable. It is true that interest

[189] *See* Samantha L. Moore-Berg et al., *Exaggerated Meta-Perceptions Predict Intergroup Hostility Between American Political Partisans*, 117 PROC. NAT'L ACAD. SCI. 14864, 14871 (2020) (concluding that "the degree to which both parties think the other side dislikes and dehumanizes their own group is dramatically overestimated.").

[190] *See* James Fishkin et al., *Is Deliberation an Antidote to Extreme Partisan Polarization? Reflections on "America in One Room,"* 115 AM. POL. SCI. REV. 1464 (2021); Magdalena Wojcieszak & Benjamin R. Warner, *Can Interparty Contact Reduce Affective Polarization? A Systematic Test of Different Forms of Intergroup Contact*, 37 POL. COMM. 789 (2020); MATTHEW S. LEVENDUSKY & DOMINIK A. STECULA, WE NEED TO TALK: HOW CROSS-PARTY DIALOGUE REDUCES AFFECTIVE POLARIZATION (James N. Druckman ed., 2021).

[191] Törnberg, *supra* note 128 at 10.

[192] A mixed-methods study that combined interviews and a survey reported results suggesting "a hunger for hard conversations" and that anonymity was valued by users (with one participant quoted as saying "when you join a social network, . . . you're exposed and you have to watch the things you write, because they can be used against you) in facilitating those (although one participant said Reddit gave them "the feeling that I am arguing over nothing with nobodies"). See Amanda Baughan et al., *Someone Is Wrong on the Internet: Having Hard Conversations in Online Spaces*, 5 PROC. ACM CONF. ON HUM.–COMPUT. INTERACTION, 8–9 (2021).

[193] See Joshua Tucker, *Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing. By Chris Bail*, 127 AM. J. SOC. 1685, 1687 (2022) ("I wonder about the extent to which DiscussIt-like platforms are really an alternative to social media platforms.").

and available time to invest in civic-minded exercises such as DiscussIt or r/ChangeMyView should not be presumed to be universal.[194] Still, they are valuable in creating an alternative environment where people can have sincere conversations about topics which have been battlegrounds of the political divide. More importantly, both r/ChangeMyView and DiscussIt suggest an exciting path for re-engineering platforms, tweaking the levers to steer the digital environment toward democracy-empowering settings, and helping to allay the illnesses afflicting politics, instead of exacerbating them. Importantly, both highlight "how the design of our platforms shapes the types of identities we create and the social status we seek."[195] Tinkering with anonymity and identification as such is an important component of potential experimentation with other social media affordances and should be a focus of attention when considering the other kinds of internet ecosystems that scholars such as Ethan Zuckerman have imagined.[196]

Conclusion

The plurality of identification and plurality of anonymity emerging from the study of networked communities holds cross-cutting insights. This paper begins the work of setting out what it entails. Identity is not fixed but is perennially shaped by and shapes group identity and norms, as Robert Post explicates.[197] Real names as used in a networked society are not equivalent to how real names work offline. Anonymity has been wrongly conceived as a marker of the absence of communal identity and of community norms. In effect, it is an ingredient in *establishing communities*,[198] mediated by other affordances, including design, norms, and practices.

[194] *See* Bail, *supra* note 113 at 132: "Needless to say, not everyone would use a platform where you gain status for bridging political divides."

[195] *Id.* at 128.

[196] *See* Ethan Zuckerman, *The Case for the Digital Public Infrastructure*, 20-01 Knight First Amend. Inst. (Jan. 17, 2020), https://knightcolumbia.org/content/the-case-for-digital-public-infrastructure.

[197] *See* Robert C. Post, Constitutional Domains: Democracy, Community, Management 182 (1995): "We can define community, therefore, as a form of social organization that strives to establish an essential reciprocity between individual and social identity. Both are instantiated in social norms that are initially transmitted through processes of primary socialization and are thereafter continually reaffirmed through the transactions of everyday life."

[198] *See* Tushnet, *supra* note 19 at 108. *See also supra* note 103 and accompanying text.

We should not understand anonymity as operating according to a uniform function. Identification is likewise mediated. This point has overlooked implications for the condition of the digital public sphere revealing identification as a driver of political polarization and misinformation in a complicated interconnected machinery. Rather than a piece in that machinery, anonymity may instead afford a disentangling device to respond to the pathologies of political discourse that have concerned commentators.