Shallow Fakes

Albertina Antognini & Andrew Keane Woods*

ABSTRACT

Scholars and policymakers are rightly concerned with online deception, especially intentional efforts to spread fake news. But the problem of deception on social media is both subtler and more endemic than a series of malicious actors disseminating deepfakes. While platforms are indeed awash in fakery, many of these fakes are shallow—superficial tweaks to one's self-presentation—and they are of the platforms' own making.

Every day on social media, users place filters on their selfies, post photos out of context, and otherwise present a fake version of their lives. This is no accident. The ability to curate a better-than-real image is the sine qua non of social media platforms, whose business model relies on blurring the distinction between true and false, authentic and inauthentic, and, ultimately, content and advertising.

We argue that this widespread superficial fakery leads to a host of underappreciated costs. At a bare minimum, the sheer scale of deception warrants greater scrutiny, which would require more information sharing from the platforms. What little we do know is troubling. The platforms' internal research shows that users, especially younger users, feel enormous pressure to adhere to a specific ideal of beauty. The pressure to conform is intense and manifests itself in traditionally gendered and racialized ways, with harms often falling on already-marginalized groups. Then there are epistemic and democratic concerns. The erosion of public trust and political polarization are often pinned on digital echo chambers, foreign influence campaigns, or both. But what share of the blame belongs to the

* James E. Rogers Professor of Law and Milton O. Riepe Professor of Law, respectively, University of Arizona James E. Rogers College of Law. The authors thank workshop participants at the Technologies of Deception conference organized by Yale Law School's Information Society Project, Stanford Law School's Grey Fellow's Forum, the Family Law Scholars and Teachers Conference, and the University of Arizona. We are especially grateful for feedback from Michael Boucai, Andy Coan, Derek Bambauer, Jane Bambauer, Ellie Bublick, Joanna Grossman, Aníbal Rosario-Lebrón, George Fisher, Jill Hasday, Lynne Henderson, Xiaoqian Hu, Mugambi Jouet, Shalev Roisman, Charisa Kiyô Smith, Tyler Valeska, Deepa Varadarajan, and Tammi Walker.

fact that so much of everyday life takes place in a space that is marked by constant, casual deception?

This Article defines shallow fakes and explains their centrality to the social media ecosystem. It then turns normative, assessing the costs of shallow fakes, which often slip through the hard and soft law that govern other kinds of public information sharing, like advertising and journalism. We end with prescriptions, chief among them a need for more transparency around how the platforms operate.

Table of Contents

I. INTRODUCTION

Social media is awash in fakery.[1] Scholars and policymakers have become especially worried about malicious disinformation tools like "deepfakes"—hyper-realistic fake videos made with artificial intelligence.[2] But the problem of deception online is both subtler and more endemic than a series of bad actors engaged in information warfare. Most of today's online fakes are actually quite shallow—superficial tweaks to one's self-presentation.[3] Every day on social media, users place filters on their selfies, post photos out of context, and otherwise present a digitally-enhanced version of their lives. Unlike deepfakes, these superficial tweaks to one's self-presentation—which we term "shallow fakes"—are enabled and encouraged by the platforms.

The ability to curate a better-than-real image is the *sine qua non* of social media platforms. Instagram, for example, owes its start to the filter:

1. *See, e.g.*, Suroush Vousoughi et al., *The Spread of True and False News Online*, 359 SCIENCE 1146, 1148–49 (2018) (examining 12 years of Twitter data and showing that "[f]alsehood reached more people" than the truth and that users were 70% more likely to share fake news than real news); CAILIN O'CONNOR & JAMES OWEN WEATHERALL, THE MISINFORMATION AGE: HOW FALSE BELIEFS SPREAD 147–67 (2019) (explaining how social networks are particularly fertile breeding grounds for misinformation); RICHARD L. HASEN, CHEAP SPEECH: HOW DISINFORMATION POISONS OUR POLITICS—AND HOW TO CURE IT (2022) (explaining the particular impact social media networks and fake news can have on democratic elections and proposing legal reforms); Peter Sucio, *Social Media Is Full of Fakes—As In Fake Followers New Study Finds*, FORBES (Nov. 17, 2021, 1:09 PM), https://bit.ly/3ooJbUB (describing how over a third of top influencer's followers are fake accounts); Hunt Allcott & Matthew Gentzkow, *Social Media and Fake News in the 2016 Election*, 31 J. ECON. PERSPECTIVES 211, 219–23 (2017) (discussing data that shows the pervasiveness of fake news on social media in the leadup to the 2016 presidential election).

2. *See* Bobby Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CALIF. L. REV. 1753, 1757 (2019) (describing the problem of "[t]echnologies for altering images, video, or audio . . . in ways that are highly-realistic and difficult to detect"). There are many different definitions of deepfakes, but they all emphasize the use of sophisticated technology, specifically artificial intelligence (AI) and deep learning, to manipulate content and deceive consumers. Indeed, the word "deepfake" is a "portmanteau of 'deep learning'"—a reference to a kind of AI algorithm—and "fakes." James Vincent, *Why we need a better definition of 'deepfake'*, THE VERGE (May 22, 2018, 2:53 PM), https://bit.ly/3WltoCB. Deepfakes originated in pornography, when, in 2017, one anonymous Reddit user uploaded pornographic videos featuring celebrities; most of the celebrities were female. *See* Russell Spivak, *"Deepfakes": The Newest Way to Commit One of the Oldest Crimes*, 3 GEO. L. TECH. REV. 339, 345–46 (2019). For one example among many of how advocates are responding to the deepfake problem, see *Prepare, Don't Panic: Synthetic Media and Deepfakes*, WITNESS MEDIA LAB, https://bit.ly/43sz1Sj (last visited June 24, 2023).

3. This term is distinct from "Shallowfakes," which is defined as "videos that have been manipulated with basic editing tools or intentionally placed out of context." HENRY AJDER ET AL., THE STATE OF DEEPFAKES: LANDSCAPE, THREATS, AND IMPACT 11 (2019), https://bit.ly/3BH6Se4. Both terms address low-technology edits, but that is where their similarities end. We are not concerned with whether the user intends to deceive or whether it is malicious. Our use of "shallow" is meant to both imply the surface level changes we focus on along with their perceived unimportance.

at a time when photo-sharing tools had become commonplace, the application stood out for its image-enhancing filters, which gave photos an attractive, sepia-tinted look.[4] The popularity of Instagram's filters set off an arms race for digital beautification.[5] Snapchat introduced video filters, which aggressively alter the appearance of one's face.[6] Even the most playful Snapchat filters, like those that add Harry Potter glasses or a crown of flowers, change users' skin color, jawline, eye shape, and more.[7] TikTok also offers its own filters. The race to beautify images is so intense that TikTok users report finding beautification filters applied to their images even when they have selected "no filter."[8]

This dynamic is largely a result of platforms' competition to attract users. The business model of today's biggest social media platforms[9] depends on bringing users to their platforms to monetize their attention through advertisements.[10] Users are exposed to multiple layers of

4. *See* SARAH FRIER, NO FILTER: THE INSIDE STORY OF INSTAGRAM xxi (2020) ("[B]ecause of filters that initially improved our subpar mobile photography, Instagram started out as a place for enhanced images of people's lives. Users began to accept, by default, that everything they were seeing had been edited to look better."); *see also* Amelia Tait, *How Instagram changed our world*, THE GUARDIAN (May 3, 2020, 6:00 AM), https://bit.ly/3Olir20 ("When Instagram launched, it offered filters that people could use to make their photos—and by extension, their lives—look more appealing.").

5. In some ways, Snapchat's early success can be attributed to the playful filters that allowed users, but especially young users, a way to be silly and rebel from the pressure to conform to the kind of look one found on Instagram; eventually, though, that would change. *See* FRIER, *supra* note 4, at 114, 179–207.

6. *See id.* at 113.

7. *See* Andrea Navarro, *Snapchat's "Pretty" Filters Allegedly Make You Whiter*, TEEN VOGUE (May 16, 2016), https://bit.ly/41R0Fqu (describing how filters like "flower crown" do more than they first appear, including whitening and smoothing skin, thinning the jawline, and more).

8. *See* Abby Ohlheiser, *TikTok changed the shape of some people's faces without asking*, MIT TECH. REV. (June 10, 2021), https://bit.ly/43aQSwn (describing users discovering that TikTok applied beauty filters to users who had all filters turned off). Additionally, Tristan Harris, the Executive Director for Human Technology, states:

> Unless the government acts, the competition between technology businesses' never-ending interest in capturing human attention, will irreversibly dismantle the information environment, accelerate polarization leading towards civil war, degrade the mental health of a generation of children and teenagers, and break down the basis for trust itself, leading to market collapse and near permanent civil disorder.

Optimizing for Engagement: Understanding the Use of Persuasive Technology on Internet Platforms: Hearing Before the Subcomm. on Commc'n, Tech., Innovation, and the Internet, of the S. Comm. on Commerce, 116 CONG. 50, 58 (2019) (Statement of Tristan Harris, Exec. Dir. Ctr. for Humane Tech.).

9. We are addressing here the advertising-driven social media platforms—typified by Facebook, Instagram, YouTube, TikTok, and Snapchat. There are smaller platforms, like many dating apps, that are not advertising-driven.

10. *See* TIM WU, ATTENTION MERCHANTS: THE EPIC SCRAMBLE TO GET INSIDE OUR HEADS 5 (2016) ("Over the last century, . . . we have come to accept a very different way

advertising at once, with varying degrees of transparency. Companies pay platforms like Instagram to present products in their users' feeds.[11] These ads can be targeted, based on user preferences, and are typically marked as advertisements.[12] There are also influencers—social media users who make a living depicting a particular lifestyle—who sell products and services they are paid by companies to promote.[13] These promotions are not always identifiable as advertisements. One study found that though nearly a quarter of all Instagram posts were advertisements, most of those had no label signifying as much.[14] Another study found that 93% of sponsored content by top influencers did not comply with Federal Trade Commission (FTC) guidelines for endorsements.[15] This is not accidental; softening the line between authentic and inauthentic is the reason native advertising works.[16] Even the initial invitation from social media platforms to users traffics in this muddling of boundaries by getting users to spend time in an advertising setting that does not feel like one.[17]

The muddling goes deeper than ambiguous advertisements. The platforms have created a market where *all* users, influencers or not,

---

of being, whereby nearly every bit of our lives is commercially exploited to the extent it can be.").

11. *See* Amanda Reaume, *How Does Instagram Make Money for Facebook (Meta Platforms)*, SEEKING ALPHA (Dec. 1, 2021, 9:00 AM), https://bit.ly/3pVgKhE.

12. *See id.* (describing the different ways that ads can be presented to users).

13. As one recent description explains:

> Social media influencers are people with extensive social media followings who share content on Instagram, TikTok, Twitter, Facebook, and other social media applications . . . . Influencers receive money from brands to promote various products to their followers. An influencer's ability to earn money from promotions correlates with their number of followers.

Stasia Skalbania, Comment, *Advising 101 For the Growing Field of Social Media Influencers*, 97 WASH. L. REV. 667, 669–70 (2022) (identifying four categories of followers based on the number of followers, which begin at "nano" and end in "mega" influencers).

14. *See Influencer Ad Disclosure on Social Media: A Report Into Influencers' Rate of Compliance of Ad Disclosure on Instagram*, ADVERTISING STANDARDS AUTHORITY REPORT 4 (Mar. 18, 2021), https://bit.ly/43eeb8A.

15. *See 93% of Top Celebrity Social Media Endorsements Violate FTC Guidelines*, MEDIAKIX (May 31, 2017), https://bit.ly/3Q9o4kF.

16. *See* Lili Levi, *A "Faustian Pact"? Native Advertising and the Future of the Press*, 57 ARIZ. L. REV. 647, 665 (2015) (arguing that the "entire *raison d'être* [of native advertising] is precisely to disable consumers from being able to distinguish between editorial content and commercial propaganda—to trick consumers and end-run ad avoidance"); *see also* FRIER, *supra* note 4, at 138 (describing the power of the unlabeled paid post on Instagram: "[s]ince consumers are much more likely to be swayed to buy something if friends or family recommend it, as opposed to advertisements or product reviews, these ambiguous paid posts were effective").

17. Take Facebook's stated goal: "to give people the power to build community and bring the world closer together." Andy Wu, *The Facebook Trap*, Technology and Analytics, HARVARD BUSINESS REVIEW (Oct. 19, 2021), https://bit.ly/3oi3QK1. Of course, that is not the only goal. Connecting users and increasing their engagement with each other "directly drive advertising revenue, the predominant mode by which Facebook captures value, i.e., monetizes the user base that otherwise uses Facebook for free." *Id.*

compete for attention and for "likes" by posting filtered, edited, or otherwise enhanced images of themselves.[18] A recent survey found that 90% of women regularly apply filters to their selfie photos.[19] One of the most popular apps ever developed—a top-downloaded app for five years in a row, now among the handful of billion-dollar unicorn apps—is FaceTune, which allows users to digitally alter their photos.[20] As a result of this casual deception, users spend hours every day in a world where most of the images they see are manipulated. From the platform's perspective, this is a feature, not a bug.[21]

To date, this widespread, superficial fakery has largely gone unnoticed. Perhaps that is because it seems at best empowering and at worst harmless. If users want to tweak their appearance online, that is their prerogative. One could argue that filters are an autonomy-enhancing form of digital identity.[22] Under this account, filters allow for a positive form of play, self-expression, or self-discovery.[23] Moreover, today's social media users are savvy—they know that what happens online is not real, so no one is being deceived or harmed. Even if this fakery did lead to some discernable harm, this kind of subtle deception is certainly not new: cropping, photoshopping, and airbrushing are old techniques.[24] Under any of these accounts, shallow fakes do not merit serious consideration.

---

18. *See* FRIER, *supra* note 4, at xxi ("Users began to accept, by default, that everything they were seeing had been edited to look better. Reality didn't matter as much as aspiration and creativity.").

19. *See* Sarah Fielding, *90% of Women Report Using a Filter on Their Photos*, VERYWELL MIND (Mar. 15, 2021), https://bit.ly/3MFN7Zx.

20. *See* Connie Loizos, *The maker of popular selfie app FaceTune just landed $135 million at unicorn valuation*, TECHCRUNCH (July 31, 2019, 7:00 AM), https://bit.ly/3Co4CII.

21. *See* Alexandra J. Roberts, *False Influencing*, 109 GEO. L.J. 81, 84 (2020) (describing how "[a]uthenticity lies at the core of the [influencer] advertising model" which "creates an exceptionally fertile breeding ground for deception and consumer harm"); *see also* Georgia Wells et al., *Facebook Knows Instagram is Toxic for Teen Girls, Company Documents Show*, WALL ST. J. (Sept. 14, 2021, 7:59 AM), https://on.wsj.com/3MHDfQ7 ("The features that Instagram identifies as most harmful to teens [things like trying to live a perfect life online, having a perfect body, only sharing one's best moments] appear to be at the platform's core.").

22. *See* Sofia P. Caldeira et al., *Exploring the Politics of Gender Representation on Instagram: Self-representations of Femininity*, 5 DIGEST. J. DIVERSITY & GENDER STUD. 23, 25 (2018) ("[T]here is still a sense of optimism surrounding the political potential of self-representation on apps such as Instagram.").

23. *See*, *e.g.*, JULIE E. COHEN, CONFIGURING THE NETWORKED SELF: LAW, CODE, AND THE PLAY OF EVERYDAY PRACTICE 53–58 (2012) (arguing for "a renewed appreciation for the play of everyday practice" and the centrality of the idea of "play" to mature conceptions of the networked self and information law).

24. *See generally* HANY FARID, FAKE PHOTOS (2019) (describing a long history of photographic manipulation, including by leaders like Stalin and Mussolini, who regularly doctored photographs to achieve political ends).

We disagree. At a bare minimum, the scale of visual deception described here—and its relationship to disinformation and the overall health of the digital ecosystem—warrants greater attention. Much of what we know about the platforms comes from whistleblowers, like former Facebook employee Frances Haugen, whose leaks to the Wall Street Journal led to international outcry and a round of Congressional hearings.[25] Knowing more about the impact of social media—and the casual deception it engenders—would require drastically more transparency from the platforms.

What little we do know is troubling. The extent and type of shallow fakes documented here suggest that users are under enormous pressure to conform to standards dictated by the platforms. If users were fully in control, we would expect filters to amplify the diversity of the online visual field, making people as distinctly unique as they want to be. But that is not what we find. Instead, millions of teens use the same filter that whitens their skin to the same hue and complain that they feel drained by the constant need to comply with a particular beauty standard.[26] This beauty standard is not simply an organic reflection of community values; rather, it is the result of their interactions with the platforms' algorithms.[27] That this dynamic exists and that it is harmful is well known to these platforms. Facebook's own internal research suggests that time spent on Facebook and Instagram has a profound negative effect on the mental health of its users, leading to anxiety and depression.[28] Independent research corroborates these findings, noting that a significant number of suicidal teens said their darkest thoughts were prompted by the platform and that the firm's products made "body image issues worse for one in

25. *See the facebook files*, WALL ST. J. (2021), https://bit.ly/3PiJeg1 (last visited June 24, 2023) (collecting news reports from the fall of 2021 that rely on leaked internal Facebook documents to describe that "platforms are riddled with flaws that cause harm, often in ways only the company fully understands"); *see* Reed Albergotti, *Frances Haugen took thousands of Facebook documents: This is how she did it,* WASH. POST (Oct. 26, 2021, 12:00 PM), https://bit.ly/3rIqGvS ("For nearly a month, Haugen has made headlines for her decision to blow the whistle on Facebook, testifying in front of Congress, appearing on "60 Minutes" and on the cover of Time Magazine. Her revelations have created a firestorm. And Facebook is reportedly considering a name change.").

26. *See* Rosalind Gill, *Changing the perfect picture: Smartphones, social media and appearance pressures*, CITY UNIVERSITY OF LONDON 5 (2020), https://bit.ly/3q01Mr4.

27. *See infra* Sections III.A and III.B.

28. *See* Wells et al., *supra* note 21. There were counter-studies—the internal research was not all negative—but there was enough information that was worrying. *See* Instagram Press Release, *What Our Research Really Says About Teen Well-Being and Instagram* (Sept. 26, 2021), https://bit.ly/41Tk9ef.

three teen girls."[29] The pressure to conform to traditional norms can also exclude nonbinary users.[30]

The social media ecosystem is not only deeply gendered, it is also racialized. Reports of blackfishing are common, in which white users go to extreme lengths, including darkening their skin, to appear Black.[31] There are also simultaneous reports of whitewashing, in which filters whiten the skin of non-white users and perpetuate various forms of digital exclusion.[32] With race, as with gender, digital tools appear to cheapen and flatten user diversity.

Then there are epistemic and democratic concerns. What happens to truth in a world where so much of everyday life is marked by constant deception? And what are the implications for democracy and public discourse? As people spend more time in online spaces where deception is the norm, what happens to democratic deliberation? Blame for the erosion of public trust and political polarization is often pinned on digital echo chambers, foreign influence campaigns, or both.[33] But we propose that some share of the blame belongs to the fact that so much of everyday life takes place in a space that is marked by sustained, but subtle, deception.[34]

---

29. *See* Instagram Press Release*, supra* note 28.

30. *See infra* Section IV.A.5; *see also* Hattie Garlick, *Why gender stereotypes are perpetuated on Instagram*, FINANCIAL TIMES (Mar. 13, 2020), https://bit.ly/44RA9PA ("[I]n the strange world of social media, new pressures are perpetuating antiquated gender stereotypes."). *But see* Jenna Wortham, *On Instagram, Seeing Between The (Gender) Lines*, N.Y. TIMES (Nov. 15, 2018), https://nyti.ms/3r2f1HV (arguing that social media "has turned out to be the perfect tool for nonbinary people to find and model their unique places on the gender spectrum"). Wortham identifies the tangible benefits of community and belonging that social media provides nonbinary users, while also noting that such platforms can still be sites of social control and hostility. *Id.*

31. *See* Wesley E. Stevens, *Blackfishing on Instagram: Influencing and the Commodification of Black Urban Aesthetics*, SOCIAL MEDIA + SOC'Y 1 (Aug. 13, 2021), https://bit.ly/3of5YCe (defining blackfishing as "a practice in which cultural and economic agents appropriate Black culture and urban aesthetics in an effort to capitalize on Black markets"); *see also infra* Section IV.B.1.

32. *See* Rachel Jacoby Zoldan, *FaceApp Creator Apologizes for Whitewashing "Hot" Filter*, TEEN VOGUE (Apr. 25, 2017), https://bit.ly/439H1H7 (noting the outcry and describing how the app lightens skin tones); Neha Prakash, *Snapchat faces an outcry against 'whitewashing filters'*, MASHABLE (May 16, 2016), https://bit.ly/3WniDQb; *see also infra* Section V.B.2.

33. *See, e.g.,* ELI PARISER, THE FILTER BUBBLE: HOW THE NEW PERSONALIZED WEB IS CHANGING WHAT WE READ AND HOW WE THINK (2011) (describing how filters allow internet users to self-select into information that confirms pre-existing biases, enables radicalization, and generally erodes civil discourse); PHILIP N. HOWARD ET AL., UNIV. OF OXFORD, THE IRA, SOCIAL MEDIA AND POLITICAL POLARIZATION IN THE UNITED STATES, 2012-2018 (2018), https://bit.ly/43dh4Xc.

34. The best example we have come across that links obsession over improving one's self to misinformation about society is a piece of excellent journalism. *See* Molly Young, *How Amanda Chantal Bacon Perfected the Celebrity Wellness Business*, N.Y. TIMES MAGAZINE (May 25, 2017), https://nyti.ms/3Wnnq47 ("We tend to think of 'wellness' as

This Article is the first to fully consider the effects of these shallow fakes. There is little scholarship on this type of deception. That, we suspect, has to do with the fact that the problem is "shallow"—dealing with surface-level aesthetics and thus seen as superficial and frivolous—not to mention that much of the harm is felt by already-marginalized communities. The closest analog is a small literature on influencer marketing.[35] This gap stands in stark contrast to deepfakes, where there has been a considerable scholarly and policy response.[36] In 2020, Congress passed two bills—the U.S. National Defense Authorization Act ("NDAA") and the Identifying Outputs of Generative Adversarial Networks ("IOGAN") Act—with provisions aimed at addressing the deepfake problem.[37] Social media platforms have also responded. In the last three years, Twitter,[38] Facebook, TikTok, Snapchat, and YouTube have updated their terms of service to explicitly address deepfakes.[39] Deepfakes are undoubtedly a serious problem. But focusing solely on deepfakes—which are now banned under most platforms' inauthentic content policies—provides the false sense that what remains is authentic.[40]

We aim to remedy the gap in scholarship and policy, which is especially notable given that there are several ways the law could

---

the province of swoony liberal elites, but it does, in fact, blossom at both cultural poles."). A similar point has also been made in discussing the specific practice of "stealth marketing," in which Ellen Goodman argues that it "harms . . . by degrading public discourse and undermining the public's trust in mediated communication." Ellen P. Goodman, *Stealth Marketing and Editorial Integrity*, 85 TEX. L. REV. 83, 87 (2006).

35. *See, e.g.*, Roberts, *supra* note 21; Skalbania, *supra* note 13, at 669–70; Annamarie White Carty, *Cancelled: Morality Clauses in Influencer Era*, 26 LEWIS & CLARK L. REV. 565 (2022); Megan K. Bannigan & Beth Shane, *Towards Truth in Influencing: Risks and Rewards of Disclosing Influencer Marketing in the Fashion Industry*, 64 N.Y.L. SCH. L. REV. 247 (2019/2020).

36. *See* Chesney & Citron, *supra* note 2.

37. *See* William M. (Mac) Thornberry National Defense Authorization Act for Fiscal Year 2021, Pub. L. No. 116-283, 133 Stat. 3388 (2021); Identifying Outputs of Generative Adversarial Networks (IOGAN) Act, Pub. L. No. 116-258, 134. Stat. 1150 (2020) (to be codified in scattered sections of the U.S. Code). The 2021 NDAA was passed over the President's veto, while the President signed the IOGAN Act. *See* Scott Briscoe, *U.S. Laws Address Deepfakes,* TODAY IN SECURITY (Jan. 12, 2021), https://bit.ly/3MnzjTj.

38. Although Twitter now goes by the name X, we will continue to refer to the firm as Twitter given that this is the commonly used name for the firm's service.

39. Most of these are outright bans on any manipulated content that could lead to harm. For an example, see Vanessa Pappas, *Combating misinformation and election interference on TikTok,* TIKTOK NEWSROOM (Aug. 5, 2020), https://bit.ly/3OskL7y ("Our Community Guidelines prohibit misinformation that could cause harm to our community or the larger public, including content that misleads people about elections or other civic processes, content distributed by disinformation campaigns, and health misinformation.") But rather than taking down "synthetic and misleading media," Twitter often will label Tweets "to help people understand their authenticity and to provide additional context." *Synthetic and manipulated media policy*, TWITTER, https://bit.ly/42OrNrm (last visited June 25, 2023).

40. *See infra* Section IV.C.

intervene.[41] The most obvious and most pressing need is for more transparency from the platforms.[42] There is a growing demand for national transparency legislation, and we explain how such legislation would alleviate some of the concerns raised here. Because the platforms are advertising networks, we also explain how existing rules promulgated by FTC that prohibit deception in the marketplace can apply more broadly to social media fakery.[43] Finally, there are other measures, like industry norms and multistakeholder initiatives, that could help. Indeed, multistakeholder initiatives have had success with revising social media policies in related areas, especially with regard to violent and extremist content.[44]

The Article proceeds in four parts. Parts II and III are descriptive, outlining the many ways in which platforms provide the tools for users to engage in shallow fakery. Part II provides a taxonomy of different types of shallow fakes, and Part III explains how platforms promote and encourage them, regardless of user preference. In Part IV, the Article turns normative, assessing the costs of shallow fakes, in addition to possible benefits. Part V looks ahead to implications for regulators, scholars, and, ultimately, users.

II. WHAT ARE SHALLOW FAKES?

We define shallow fakes as superficial, commonplace deceptions about one's self-presentation online. Note that this definition does not turn on intent; we are interested both in actors who intend to deceive, for whatever reason, and those who do not. This definition covers a wide range of image-enhancement techniques, including those that were available before the rise of today's digital networks, such as airbrushing. Importantly, we begin from the premise, as Erving Goffman has observed, that "life itself is a dramatically enacted thing."[45] All of us perform a kind of "presentation of the self" every day, often by attempting to enhance our image in subtle, and not-so-subtle, ways.

41. *See infra* Part IV.
42. *See infra* Section V.A.
43. *See infra* Section V.B.
44. *See infra* Section V.C.
45. ERVING GOFFMAN, THE PRESENTATION OF THE SELF IN EVERYDAY LIFE 72 (1959). Indeed, we accept that performance is part of every social interaction and that deception routinely occurs as a descriptive matter. *See id.* at 249 ("All the world is not, of course, a stage, but the crucial ways in which it isn't are not easy to specify."). In this Article, we are concerned with the deceptions in the presentation of one's self online that are offered by the various platforms.

The aim in Part II is principally descriptive.[46] This Part begins by setting out core elements of shallow fakes; it then provides some prototypical examples. Much of the data we rely on here is necessarily incomplete because it is based on user surveys, leaked internal documents, and the platforms' own policies; these are tiny snapshots into how platforms function but, by definition, are not comprehensive. We end Part II by distinguishing shallow fakes from deepfakes, which have received the lion's share of attention thus far.

A. The Core Elements

Shallow fakes consist of four core elements. First, shallow fakes are *superficial* tweaks to one's image that are made without any specific intent and might therefore be understood as harmless. Second, they are *commonplace*. Third, they take place *online*, which makes them different from what we are accustomed to offline.[47] Fourth, and finally, they affect one's *self-presentation*; they are not principally about other people or about facts at large.

1. Superficial

Shallow fakes are superficial edits to observable characteristics. They are meant to improve the user's physical appearance. For this reason, they are typically seen as innocuous. The platforms themselves describe this kind of enhancement—the use of filters, lighting, and crops—as harmless. Instagram's policy for deceptive material, for example, only applies to edits that are "beyond adjustments for clarity or quality," which leaves considerable room for image enhancement.[48] Clarifying the platform's stance on deepfakes, one platform spokesperson explained that content is regularly manipulated "often for benign reasons" and that this content is considered authentic.[49] The seemingly benign nature of this widespread media manipulation is why Instagram can describe itself as "an authentic and safe place for inspiration and expression."[50] The paradox of shallow

46. Significantly, we are not critiquing the users who deploy the techniques we describe. In Part III, we argue that the reason there is so much fakery on platforms is the result of deliberate choices by the social media platforms.

47. In addition to the "front regions" and "back regions" that Erving Goffman identified as crucial to social performance, he also identifies "the outside." GOFFMAN, *supra* note 45, at 134–35. Online, there is no "back region" or "outside"—the audience is only privy to the front regions, which raises a set of issues specific to social media.

48. Monika Bickert, *Enforcing Against Manipulated Media*, META (Jan. 6, 2020), https://bit.ly/3qZ6Kob.

49. *Id.*

50. Community Guidelines, INSTAGRAM, https://bit.ly/3piTOsJ. The description of Community Guidelines continues: "Remember to post authentic content, and don't post anything you've copied or collected from the Internet that you don't have the right to post."

fakes is that they are subtle and superficial, which makes them less suspicious and, in turn, gives them enormous reach.[51]

2. Commonplace

Making changes to one's appearance is the norm on today's social media platforms. A survey conducted in the United Kingdom found that 90% of women aged 18–30 reported using a filter before posting online photos.[52] The purpose of these filters is to physically alter one's appearance—"to even out skin tone, reshape [the] jaw or nose, shave off weight, brighten or bronze skin, and whiten teeth."[53] Unsurprisingly, the same survey found that women feel "bombarded" and "overwhelmed" by the pressure to look as good as the other filtered images they see online.[54] One person explained, "[I]t is everywhere, all the time, and social pressure to look a certain way is very real."[55]

An indication of just how commonplace digital image distortion has become is the success of photo editing apps such as FaceTune,[56] an extraordinarily popular app which claims to "effortlessly enhance every selfie."[57] Its effects are palpable: "FaceTuning your jawline [has become] the Instagram equivalent of checking your eyeliner in the bathroom of the bar."[58] Within a year of its release, FaceTune was the number-one downloaded app in the "photo and video" category on Apple's platform in 120 countries.[59] By 2018, it had spent four years as the most-downloaded paid app worldwide and across all of Apple's app categories.[60] Lightricks, the company that makes FaceTune, is valued at $1 billion dollars, having recently raised $135 million in Series C funding.[61] The app has been downloaded 180 million times.[62]

51. *See* 1 J. THOMAS MCCARTHY, MCCARTHY ON TRADEMARKS AND UNFAIR COMPETITION § 2:22 (5th ed. 2020) ("[I]f the deception is truly effective, the consumer may not even be aware of it.").

52. *See* Gill, *supra* note 26, at 36.

53. *Id.* at 35.

54. *Id.* at 28.

55. *Id.*

56. *See* Loizos, *supra* note 20 (noting that the app "empowers users to cover their gray hairs, refine their jaw lines and reshape their noses").

57. FACETUNE, https://bit.ly/42XKwA7 (last visited Sept. 12, 2023).

58. Jia Tolentino, *The Age of Instagram Face*, THE NEW YORKER (Dec. 12, 2019), https://bit.ly/3CMehJn.

59. *See* Randy Nelson, *These Apps and Games Have Spent the Most Time at No. 1 on the App Store,* SENSOR TOWER (July 2018), https://bit.ly/3CIIER8.

60. *See id.*

61. *See* Loizos, *supra* note 20.

62. *See id.*

3. Online

Another key component of our definition of shallow fakes is that they occur online. Online deception is a different, and more difficult, problem than offline deception for at least two reasons. First, online deception is harder to identify. When one sees an airbrushed billboard of a celebrity or a model, one understands at some level that the image is not a true representation of a real person. Also, the billboard is clearly understood to be an advertisement because all billboards are advertisements. Online, the space between the "real" and the "fake" is much narrower. This is both because of who is engaging in the deception and the context in which it occurs. While comparing social media filters to airbrushing in magazines, one survey respondent explained, "What's different about social media is these aren't just celebrities and supermodels, these are people you know. The feeling of 'why isn't that me' becomes even stronger and more significant."[63]

Second, the context is less clearly defined, which means that casual, online deception seeps into all aspects of one's life. A magazine is a highly-stylized product, and it is something you can pick up and put down. Even if you access a magazine digitally, it has some distance from your everyday life and you know and expect it to be an idealized version of real life. Online airbrushing, however, is something everyone—including your friends, classmates, and colleagues—does all the time. In the aggregate, we start to live in a world where it can be hard to separate fact from fiction.

4. The Self

Shallow fakes address how individuals present themselves and their lives online in a deceptive way. It is not deception about others or about the world at large. This self-focused deception can occur in many forms, but the basic idea is that rather than choosing to reflect a mere snapshot or a moment in time, people, intentionally or not, curate the images they post in a way that distorts their lives. As one of the interviewees in a study addressing online deception noted about the kinds of things people post to social media, "[T]his isn't a realistic perception of everyday life. Things go wrong but we only want other people to see the perfect bits. You can so easily make people think you lead this picture perfect life when for most people this is not the case."[64]

This is not gross deception. It is designed to be slight—shallow fakes are meant to be small enough to be believable. This is why, for example, a British study of teenage girls' use of social media found that teens feel pressure to post images that are perfect—which requires using filters—but

63. Fielding, *supra* note 19 (quoting Tess Brigham).
64. Gill, *supra* note 26, at 23.

not so perfect that the images appear wholly unrealistic. As one subject said, "I don't want my pictures to look too fake . . . . I want it to look as natural as possible, even though I'm wearing makeup. I want it to look like I haven't put a filter on."[65]

This deception in the presentation of one's self, or one's life, is what distinguishes shallow fakes from other forms of fakery, like fake news, that have received so much attention. Indeed, one of the most widely discussed frameworks for identifying and distinguishing different types of fake news does not even mention the subtle deception in self-presentation of the kind we describe here.[66]

B. Examples of Shallow Fakes

Shallow fakes are everywhere, but they are not all the same. In this Section we provide a few paradigmatic examples to fill out our definition. While we discuss each fake separately, they often appear in conjunction. We do not mean to critique any particular user for relying on these techniques, but we do argue that this fakery is harmful.[67] In Part IV, we discuss why.

1. The Filter

Perhaps the most widespread deception today is the use of the filter: a tool that digitally enhances the appearance of a person or object. Filters can be basic, like changing the light and color in a photo. But they can be more sophisticated and now increasingly rely on artificial intelligence to change the physical appearance of an image. For example, it is common for people to use filters that change their jawline, eye shape, skin tone, and more.[68] These beautification filters are effective primarily because they are

65. *Id.* at 26.

66. *See* Claire Wardle, *Fake news. It's complicated*, FIRST DRAFT NEWS (Feb. 16, 2017), https://bit.ly/3MSz0RH; *see also* FIGHTING FAKE NEWS WORKSHOP REPORT, YALE INFORMATION SOCIETY PROJECT (2017), https://bit.ly/437iXFq.

67. We would, however, like to take this opportunity to criticize the humblebrag, which is more of a misdirection than a "shallow fake," but worth calling out, nonetheless. The humblebrag is presented as a complaint about something—say, the weight of the gold medal hanging around one's neck—but the complaint is an excuse to boast about that very thing. *See generally* HARRIS WITTELS, HUMBLEBRAG: THE ART OF FALSE MODESTY (2012). Humblebrags are just one example of a larger pattern in which people explain their motivations one way when they are, in fact, another. Law professors on Twitter are prime offenders.

68. *See* Tolentino, *supra* note 58. Tolentino writes:

> Snapchat . . . has maintained its user base in large part by providing photo filters, some of which allow you to become intimately familiar with what your face would look like if it were ten per cent more conventionally attractive—if it were thinner, or had smoother skin, larger eyes, fuller lips.

Id.

subtle. As one beauty blogger said, "If done properly, it should be hard to tell you've used it."[69] They are both understated and ubiquitous.

2. The Crop

Another way that online media can deceive is by allowing a user to present an image decontextualized from any surrounding facts. Cropped out of the frame is more information, like a broader context or another perspective. Cropping photos as a means of deception has a long history.[70] Today, the cropped photo is standard fare on social media. Cropping can be done to zoom in and make an image more visible, but it can also be done to deceive, or to hide a broader setting. One classic example of the way a crop can alter one's surroundings is the sandbox masquerading as a beach.[71] In social media, it is also common to crop out ring lights, staging equipment, and other tools used to create a highly constructed, yet seemingly natural, photograph.[72] This is the reason for the increasingly popular "behind the scenes" shot, which exposes the "ridiculous reality" behind many Instagram posts.[73]

3. The Mislabel

Content can be misleading not only because it is cropped, or taken out of context, but also because it is mislabeled. For example, a filtered image accompanied by the label "#nofilter" implies that the user applied no filter when they might have. Researchers studying the use of the "#nofilter" label found that about 12% of those labeled as such had a filter applied to them.[74]

---

69. Olivia Solon, *FaceTune is conquering Instagram—but does it take airbrushing too far?*, GUARDIAN (Mar. 9, 2018, 3:01 AM), https://bit.ly/3qcnO9I.

70. Stalin, for example, famously cropped and photoshopped his political enemies out of photographs. DAVID KING, THE COMMISSAR VANISHES: THE FALSIFICATION OF PHOTOGRAPHS AND ART IN STALIN'S RUSSIA 13 (1997) ("Many photographic deletions were not the result of retouching at all but of straightforward cropping. Art departments have always cropped photographs on aesthetic grounds, but in the Soviet Union cropping was also used with political objectives in mind.").

71. *See* Olivia Devereux-Evans, *Influencer, 20, goes viral with TikTok videos showing her glamorous 'beach' shots are actually made with buckets of sand and a kiddie pool in the backyard*, DAILYMAIL ONLINE (April 24, 2022, 5:43 AM), https://bit.ly/3N0eKO1. Interestingly, the influencer's post describing her fakery, and being transparent about how she manufactured an image, is what went viral. *See id.*

72. *See* Taylor Lorenz, *The Instagram Aesthetic Is Over*, THE ATLANTIC (Apr. 23, 2019), https://bit.ly/42gAo54.

73. *See* Rachel Hosie, *An Instagram influencer shares behind-the-scenes videos showing the ridiculous reality behind her glamorous photos*, INSIDER (Oct. 15, 2019, 7:44 AM), https://bit.ly/438ta4k.

74. *See* Renee Engeln, *The #nofilter Lie*, PSYCH. TODAY (July 23, 2019), https://bit.ly/43tUhGJ. If the 12% number holds true across Instagram, then that would mean that "there are roughly 30 million phony #nofilter pics on Instagram." *Id.*

This labeling sleight of hand is consequential in the digital realm: psychological research shows that people are more likely to believe a false statement if it is accompanied by a photograph.[75] Moreover, images are much more likely to be shared by people and promoted by the algorithms that sort social media content.[76] This kind of mislabeling can thus reach many users.

4. The Product Endorsement

One of the most common forms of deception online is a paid or sponsored post that endorses a product and is not obviously labeled as an advertisement. Advertising is ubiquitous on social media platforms—they are, after all, advertising platforms.[77] Yet one customary form of advertising is not disclosed as such. Consider the following scenario: a beer brand runs an advertisement on Instagram where sponsored posts appear and promote the beer. This post is typically obvious as an advertisement and adequately identified. But the same brand could also privately pay a popular influencer to post photos of themselves enjoying that beer on Instagram.[78] These influencers' posts do not always come with a disclaimer identifying them as advertisements: one recent study by the British Advertising Standards Authority found that nearly a quarter of *all* Instagram posts were advertisements, and yet only one third of those ads were labeled as such.[79] The study looked at 24,000 Stories on Instagram

75. *See* Deryn Strange et al., *Photographs Cause False Memories for the News*, 136 ACTA PSYCHOLOGICA 90, 90–94 (Jan. 2011) (reporting the results of an experiment that found people were more likely to remember false news reports and with more confidence if those reports were accompanied by photographs).

76. Yiyi Li & Ying Xie, *Is a Picture Worth a Thousand Words? An Empirical Study of Image Content and Social Media Engagement*, 57 J. MKTG. RSCH. 1, 1 (Nov. 18, 2019), https://bit.ly/45C427W.

77. *See* Matthew Johnston, *How Does Facebook (Meta) Make Money?*, INVESTOPEDIA (Jan. 10, 2023), https://bit.ly/45u6ld7 ("Meta Platforms (META), the company that owns Facebook, primarily makes money by selling advertising space on its various social media platforms. Major competitors include Apple (AAPL), Alphabet (GOOGL) Google and YouTube, Tencent Music Entertainment Group (TME), Amazon (AMZN), and X Corp (formerly Twitter).").

78. It is difficult to find data on this topic. Model and influencer Emily Ratajkowski explains the terms of one such exchange the following way:

> A large hotel conglomerate had just opened a new luxury resort in the Maldives. The hotel cost $400 million to build . . . . The hotel group needed to generate awareness, and having me visit and tag their account and the location was valuable to them. For this kind of advertisement, I was able to make a shit ton of money just by vacationing here for five days and posting the occasional picture.

EMILY RATAJKOWSKI, MY BODY 87 (2021).

79. *See Influencer Ad Disclosure on Social Media*, *supra* note 14, at 4.

across a three-week period and found that nearly 3,800 of those were paid advertisements with no label.[80]

C. Distinguishing Deepfakes

Finally, distinguishing shallow fakes from the better-known—and more regulated—concept of deepfakes is analytically useful. The policy responses to deepfakes are also important to briefly consider in describing the phenomenon, given how they exacerbate the lack of attention paid to shallow fakes and normalize the presence of shallow fakes online.

The core definition of a deepfake is that it uses artificial intelligence to create images that trick the eye—the kind of computer-generated images that once were available only to a well-resourced movie studio.[81] Shallow fakes, however, do not turn in any meaningful way on the use of high technology.[82] As our examples above show, consumers of online content can be misled by low-tech manipulations, like showing only part of an interaction, either from one angle or from one perspective.[83]

Also, deepfakes are described as something that is done to images of another—not to oneself. The illustrations provided by Bobby Chesney and Danielle Citron in their foundational article on deepfakes generally feature a sophisticated actor using digital tools to manipulate an image of *someone else*.[84] Shallow fakery, on the other hand, is something we do to *ourselves*. Where deepfake scholars are worried about "the creation of realistic impersonations out of digital whole cloth,"[85] shallow fakes are people merely impersonating better versions of themselves. The problem then is not that someone *else* might be falsely portrayed as endorsing a product, service, idea, or politician;[86] it is that people are portraying false presentations of themselves every day.

The discussion on deepfakes further frames the problem of online deception around bad actors with an intent to deceive. But much of online

80. *See id.* at 3; *see also* Keisha Phippen, *Are You Influencing Responsibly?*, NAT'L L. REV. (Apr. 7, 2021), https://bit.ly/43v3oXT (reporting the results of a study by the British Advertising Standards Authority).

81. *See* Chesney & Citron, *supra* note 2, at 1763 ("As the volume and sophistication of publicly available deep-fake research and services increase, user-friendly tools will be developed and propagated online, allowing diffusion to reach beyond experts.").

82. Although they can involve more sophisticated technology. *See* discussion *infra* Part III.

83. One well-documented problem is when someone posts police body camera footage that is taken out of context. *See* Emmeline Taylor & Murray Lee, *The Camera Never Lies?: Police Body-Worn Cameras and Operational Discretion*, in POLICE VISIBILITY: PRIVACY, SURVEILLANCE, AND THE FALSE PROMISE OF BODY-WORN CAMERAS 80–95 (Bryce Clayton Newell ed., 2021) (describing how police-worn camera video clips are often taken out of context).

84. *See* Chesney & Citron, *supra* note 2, at 1776.

85. *Id.* at 1758.

86. *See id.* at 1774.

deception is done with mixed motives, without malice, or maybe even without any specific intent. Focusing on actors' intent overly constrains the conversation about online deception. It also places the emphasis on the individual users, rather than on the policy promoted by the platforms. Yet the users are often merely responding to the incentives created by the platform market.[87]

Following calls from scholars and policymakers, the platforms have taken deepfakes seriously. Nearly all platforms ban deepfakes, and all try to root out what they describe as "inauthentic content."[88] This response to eradicating deepfakes assumes there is a meaningful distinction between "authentic" and "inauthentic" content. Indeed, one of the leading harms associated with deepfakes is that "a skeptical public will be primed to doubt the authenticity of *real* audio and video evidence."[89] Not only does this assertion suppose that a discernible line exists between authentic and inauthentic content, but it also characterizes as authentic a digital world that is deeply inauthentic. Carving out a set of deepfakes and calling them "inauthentic content" helps to legitimize the constant flood of shallow fakes that fill our platforms daily.

Fleshing out the problem of shallow fakes, as a companion to deepfakes, will hopefully redirect the conversation and the reform efforts currently underway.

III. PLATFORMS FOR SHALLOW FAKERY

The rise of shallow fakes is not accidental—it is the result of intentional design choices by social media platforms with the goal of attracting as many users as possible, as young as possible, in order to keep them online as long as possible. Platforms facilitate the use of shallow fakes by creating an ecosystem where applying increasingly aggressive beautification filters is the norm. The platforms also rely on algorithms and other dark patterns that leave the user with little control over the content they are exposed to online. Finally, platforms maintain a set of policies that assume there is a definitive line between "inauthentic" and "authentic" content on the platform, thereby making shallow fakes difficult to identify and address as a distinct phenomenon.

87. *See* discussion *infra* Part III.

88. *See, e.g.*, *Inauthentic Behavior*, FACEBOOK TRANSPARENCY CENTER https://bit.ly/43XtUK8 (last visited June 25, 2023) ("In line with our commitment to authenticity, we do not allow people to misrepresent themselves on Facebook, use fake accounts, artificially boost the popularity of content or engage in behaviors designed to enable other violations under our Community Standards.").

89. Chesney & Citron, *supra* note 2, at 1785 (emphasis added).

A. The Arms Race

Shallow fakes start with one person tweaking a photo, followed by another person doing so to compete with the "perfection" of the first, and then another, until the Internet is awash in unreal and unrealistic images. At least two different kinds of "arms races"[90] lead to this cascade of fakery. The first race is among individual users. The second is between the various platforms.

In a world where people vie for attention and swipes, if one person uses beauty filters to enhance their image, the incentive is for everyone else to do the same.[91] The effect goes well beyond influencers; indeed, all "users know they will be judged" by the "followers, likes, and comments" they seek for various reasons, including "personal validation, social standing, and even financial reward."[92] It is against this backdrop that filters help "to ease the pressure of gaining likes and followers."[93] And the pressure does build. As one study participant put it, "[E]ven when you say, oh, I'm fine, that doesn't bother me, in the back of your mind, it is still like, well, everyone else on Instagram looks this way."[94] Data on teen users, in particular, show that teens take "dozens of different angles of the same shot, finding the perfect one, then edit[] away their imperfections before posting."[95] If their photo does not get enough likes, they delete it.[96]

This user activity does not occur in a vacuum.[97] It is a direct result of the second arms race, which is taking place among the platforms—to have people use, and remain on, their sites. What that means for each platform varies. The purpose of the Instagram filter, which "give[s] Instagrammers

90. Tristan Harris used the term "arms race" to describe the pressure the platforms experience to "beautify" their users. *See* ELISE HU, FLAWLESS: LESSONS IN LOOKS AND CULTURE FROM THE K-BEAUTY CAPITAL 132 (2023) (quoting Tristan Harris).

91. In chronicling the rise of Instagram, Sarah Frier notes how it affected everyone on the platform, becoming "a tool for crafting and capitalizing on a public image, not just for famous figures but for everybody." FRIER, *supra* note 4, at 128.

92. *Id.* at 233.

93. *Id.* at 173.

94. Gill, *supra* note 26, at 30.

95. FRIER, *supra* note 4, at 114.

96. *See id.* (explaining that "they would often delete pictures if they didn't get 11 likes," which "was the number of likes that would turn a list of names below an Instagram post into a number—a space-conserving design that had turned into a popularity tipping point for young people"). In a hall-of-mirrors sort of way, teens would have separate accounts called "finstas," short for "fake Instagram" accounts where they could post images that were more realistic and unedited. *Id.* at 182–83. These accounts were, for the most part, private. *Id.* This is also the reason for the widespread meme known as "felt cute, might delete." *See Feeling Cute, Might Delete Later*, KNOW YOUR MEME, https://bit.ly/3sRUHda (last visited Sept. 12, 2023).

97. *See, e.g.*, FRIER, *supra* note 4, at 278–79 ("Instagram isn't designed to be a neutral technology, like electricity or computer code. It's an intentionally crafted experience, with an impact on its users that is not inevitable, but is the product of a series of choices by its makers about how to shape behavior.").

permission to present their reality as more interesting and beautiful than it actually was" is also "exactly what would help make the product popular."[98] Instagram rewards likes, comments, and followers.[99] The platform gives users nearly immediate feedback on their posts, and the images that repeatedly lead to the "best" results according to Instagram's metrics are "airbrushed selfies, crazy action shots, and scantily clad influencers."[100] YouTube, on the other hand, rewards creators for how much time the user spends watching their content.[101] YouTube measures that time by considering the percentage of a video viewed and the average duration that the user watched.[102] To succeed according to YouTube's metrics then, creators have to alter their format and transition to longer videos, like the "15-minute makeup tutorial videos."[103] They also must tailor their material to keep viewers watching, which might mean presenting off-the-wall conspiracy theories.[104]

All social media platforms share the need to grow their user-bases, and this enters them into an arms race amongst themselves to do just that. This has meant that each company attempts to out-offer the other; it has become an inevitable feature of the marketplace for customers. In the context of shallow fakes, if one company employs a beauty filter that attracts customers, that will cause the other companies to develop something similar.[105] This competition explains why Snapchat, which initially began as an antidote to those apps that "conform to unrealistic notions of beauty or perfection" and aimed to create a space for users "to be funny, honest, or whatever else you might feel like at the moment,"[106] spent $150 million to buy a company, Looksery, that uses facial recognition technology to "photoshop video chats and messages in real time."[107] It explains why Facebook, after failing to acquire Snapchat, bought the Masquerade app, which allows people to digitally enhance their

98. *Id.* at 23.
99. *See id.* at 234.
100. *Id.* at 238.
101. *See id.* at 233.
102. *See id.*
103. *Id.*
104. *See id.*
105. As Tristan Harris has said, each company competes to provide resources, like beauty filters, that the other does not have. *See* #1736 - Tristan Harris & Daniel Schmachtenberger, THE JOE ROGAN EXPERIENCE, SPOTIFY (Nov. 2021), https://spoti.fi/3JwPBs2 ("[I]f one attention company doesn't add the beautification filter, the other one will.").
106. Sophia Bernazzani, *A Brief History of Snapchat*, HUBSPOT (Oct. 29, 2019), https://bit.ly/3MTu1QD (quoting Snapchat founder Evan Spiegel).
107. Alyson Shontell, *Snapchat buys Looksery, a 2-year-old startup that lets you Photoshop your face while you video chat*, BUSINESS INSIDER (Sept. 15, 2015, 3:36 PM), https://bit.ly/3MF5H3Y.

selfies with filters like Snapchat's.[108] And it explains why the beautification app FaceTune is allegedly worth more than a billion dollars.[109] This race-to-the-bottom dynamic also explains why the platforms have been caught applying slight beautification filters by default without warning users, even where the user has not selected a filter.[110]

B. Deception in the Algorithm

Social media platforms not only provide a medium where fakery by users is the norm, but they also employ dark patterns, which are user interfaces "whose designers knowingly confuse users, make it difficult for users to express their actual preferences, or manipulate users into taking certain actions."[111] The premise of the algorithms is to show a user content they will like, presumably based on the user's past actions on the platform.[112] Users are accustomed to a feed of information that regularly suggests the next article, the next link, the next video. That feed, however, does not just reflect user preferences. Because the feed is optimized to, above all, keep a user engaged, it also manipulates those preferences, nudging the viewer in one direction or another. Suppose a user searches YouTube for "cooking videos." The results might also recommend a video of a lion wrestling with a crocodile, if it thinks that user will watch it. In this way, the algorithm has deceived the user—it purported to show *relevant* results and instead it showed results a user might click and watch just because the video is enticing. A Pew Center study confirmed this phenomenon, revealing that the YouTube recommendation system pushed users into watching "progressively longer and more popular content," regardless of what had been previously viewed.[113]

108. *See* Paresh Dave, *Facebook buys Masquerade, app company that competes with Snapchat's lenses*, L.A. TIMES (Mar. 9, 2016, 10:24 AM), https://bit.ly/43mDaH4.

109. *See* Loizos, *supra* note 20.

110. *See* Ohlheiser, *supra* note 8.

111. *See* Jamie Luguri & Lior Jacob Strahlevitz, *Shining a Light on Dark Patterns*, 13 J. LEGAL. ANAL. 43 (2021) (reporting the results of studies that illustrate how dark patterns work).

112. *See* Clodagh O'Brien, *How Do Social Media Algorithms Work?*, DIGIT. MKT. INST. (Jan. 19, 2022), https://bit.ly/3BRGXAa ("Algorithms are used on social media to sort content in a user's feed. With so much content available, it's a way for social networks to prioritize content they think a user will like based on a number of factors.").

113. Aaron Smith et al., *Many Turn to YouTube for Children's Content, News, How-To Lessons*, PEW RESEARCH CENTER (Nov. 7, 2018), https://bit.ly/3MBz2vU. One of the findings shows that 28% of the "unique videos" recommended to users "were recommended more than once over the study period, suggesting that the recommendation algorithm points viewers to a consistent set of videos with some regularity. In fact, a small number of these videos (134 in total) were recommended more than 100 times." *Id.* Similarly, "regardless of whether the initial video was chosen based on date posted, view count, relevance or user rating," the YouTube algorithm "consistently suggested more popular videos." *Id.*

This dynamic is precisely what makes TikTok's algorithm famously addictive. The algorithm shows users things that it knows will keep them on the platform, based only in part on the users' interests.[114] One interpretation of the algorithm is that it seems to know you better than you know yourself.[115] Perhaps. But revealing one's true self is not the algorithm's purpose; its purpose is, above all else, to keep users on the site. In so doing, the algorithm ends up showing users a specific slice of the internet, one that is being promoted by companies rather than by the will of the individual user.

Take the example of on-line pornography: PornHub, which provides free online pornography, relies on algorithms of the same kind used by other large companies like Amazon, Netflix, and Facebook.[116] These algorithms track data to learn about their users, including their search histories, location, and times when they are online. Even by merely aggregating this data, they direct the content that users see. Thus, someone with otherwise unconventional sexual desires might find themselves pushed by the data into "very stereotyped, often sexist, often racist ways, and also just with a narrow-minded view of sexuality."[117] The result, as is true on other online platforms, is that "[o]nline-porn users don't necessarily realize that their porn-use patterns are largely molded by a corporation."[118] What is presented as the mere extension of one's preferences is, in fact, shaping those preferences, with the ultimate goal of increasing time spent on the platforms. Today's platform algorithms do not just reflect users' preferences—they also mold them.

C. Platform Policies on Deception

Platform policies have begun to address the problem of online deception. But in doing so, they have unwittingly exacerbated the problem of shallow fakes. As an initial matter, the policies are too broad to implement effectively. The policies are also inconsistent. They prohibit

114. *See* Ben Smith, *How TikTok Reads Your Mind*, N.Y. TIMES (Dec. 5, 2021), https://bit.ly/3qcm5RC.

115. *See id.* ("[T]he app is shockingly good at reading your preferences and steering you to one of its many 'sides,' whether you're interested in socialism or Excel tips or sex, conservative politics or a specific celebrity. It's astonishingly good at revealing people's desires even to themselves.").

116. *See* Joe Pinsker, *The Hidden Economics of Porn*, THE ATLANTIC (Apr. 4, 2016), https://bit.ly/3OzykSD.

117. *Id.* Shira Tarrant, author of THE PORNOGRAPHY INDUSTRY, elaborates: "If you are interested in something like double oral, and you put that into a browser, you're going to get two women giving one guy a blowjob . . . you're not likely to get two men or two people giving a woman oral sex." *Id.*

118. *Id.*; *see also* AMIA SRINIVASAN, THE RIGHT TO SEX 67–68 (2021) (describing how "free online porn doesn't just reflect preexisting sexual tastes" given the ways that companies use algorithms).

deception in some contexts only some of the time; in other instances, deception is prohibited in theory but allowed in practice. At worst, the policies allow or even openly encourage deception. Further, the policies are selectively and inconsistently enforced. Such inadequate attempts at rooting out obvious instances of deception, combined with the focus on deepfakes, entrenches shallow fakes by creating the impression that the remaining content is authentic and therefore trustworthy.

The sheer breadth of these policies conveys conflicting messages about when deception is prohibited. As a simple example, consider the following statement from Snapchat's Terms of Service: "We prohibit pretending to be someone (or something) that you're not, or attempting to deceive people about who you are."[119] That sounds like both a reasonable effort by the platform to root out deception and also a description of the majority of Snapchat posts. Pretending to be someone or something else is among the most common things that people do online.[120] Instagram provides yet another illustration. Its terms of use, which all users agree to by signing onto the platform, include: "You can't do anything unlawful, misleading, or fraudulent or for an illegal or unauthorized purpose."[121] Reading this, one might think that posting a photograph of someone posing on a towel in a sandbox, designed to make it look like a beach, would be banned. But it is not. Although obviously "misleading," it is allowed.

Perhaps this is because the platforms' own terms prohibiting deception in one context are incompatible with terms that refer to "misleading" content in other contexts. When Meta (then Facebook) announced in 2018 that it was ramping up its efforts to reduce the spread of "false information," it did not define the term. It then noted that "[a]lthough false news does not violate our Community Standards, it often violates our policies in other categories, such as spam, hate speech or fake accounts, which we remove."[122] In a different set of policies, titled "Reducing the Spread of False Information on Instagram," the platform declared that it may remove or reduce the availability of anything that is "false information, altered content, or content with missing context."[123] Read together, a piece of false information could be "misleading" under

119. *Community Guidelines*, SNAP (Aug. 2023), https://bit.ly/43bq8MY.

120. *See* Priya Singh, *This holiday, let's stop this social media pretending*, CNN (Dec. 23, 2021, 9:39 PM), https://bit.ly/3IDVraU ("Everyone from regular people to Olympic athletes and Fortune 500 CEOs feel an unending pressure to pretend everything's not just OK but is actually great. Certainly, it's compounded by the pressure to measure oneself against the distorted images of peers on social media.").

121. *Terms of Use*, INSTAGRAM, https://bit.ly/3qaxQZ7 (last visited Sept. 12, 2023).

122. Tessa Lyons, *Hard Questions: What's Facebook's Strategy for Stopping False News?*, META (May 23, 2018), https://bit.ly/3Mxaep0.

123. *Reducing the Spread of False Information on Instagram*, INSTAGRAM, https://bit.ly/3CN1eri (last visited Sept. 12, 2023).

the community standards policy, but not the fake accounts or hate speech policy, and vice versa. There is no consistent definition for what counts as "misleading."

Then there is the one-off narrowing of the type of deception that is targeted. Some deception, for example, is banned only when it is monetized, and only when it is monetized in certain ways. These rules further confuse the kind of deception that is prohibited. For example, Facebook openly acknowledges that "[a] lot of the misinformation that spreads on Facebook is financially motivated."[124] To that end, the firm's Instagram Content Monetization Policies do not allow "[m]isinformation," which they define as "content that has been rated false by a third-party fact checker."[125] But, that kind of information is implicitly allowed as long as it is not "monetized."[126] Moreover, it is presumably the case that this statement means that the *user* cannot monetize the misleading content, but Facebook can because Facebook can still drive engagement (and therefore advertising sales) to its platform by allowing, and even recommending, deceptive content by algorithm.[127]

Unsurprisingly, then, enforcement is inconsistent and unpredictable. The sweeping scope of these policies means that platforms have a great deal of discretion in deciding how to implement them. Indeed, platforms have a record of selectively intervening to manage content they deem deceptive.[128] The mechanisms the platforms rely on for identifying misinformation are partly at fault. Most of the misinformation discovered on Instagram depends on reports from users and verification from a "third-party fact checker,"[129] which limits enforcement of the misleading and deceptive content rules to a tiny range of deception on the platform. The majority of content will therefore not be flagged as deceptive—and, ironically, the more deceptive it is, the less likely it is to be flagged, given that the deception will be well-hidden. Of the content that is flagged, action will be taken only if and when such content is reviewed by a third party. Content moderation is thus limited to what can be verified as

124. Lyons, *supra* note 122.

125. *Instagram Content Monetization Policies*, INSTAGRAM, https://bit.ly/43XWOd8 (last visited Sept. 12, 2023).

126. *Id.*

127. *See* Ryan Mac & Cecilia Kang, *Whistle-Blower Says Facebook 'Chooses Profits Over Safety'*, N.Y. TIMES (Oct. 3, 2021), https://bit.ly/3prct5t ("In reality, Facebook knew its algorithms and platforms promoted this type of harmful content, and it failed to deploy internally recommended or lasting countermeasures.") (quoting Frances Haugen).

128. *See* Jeff Horwitz, *Facebook Says Its Rules Apply to All. Company Documents Reveal a Secret Elite That's Exempt*, WALL ST. J. (Sept. 13, 2021, 10:21 AM), https://bit.ly/3Pv7NX9 (describing how the firm's enforcement of its own rules is uneven, and nonexistent for a group of select users who had been whitelisted).

129. *Instagram Content Monetization Policies, supra* note 125.

misleading. If something cannot be verified as misleading, presumably cheaply and quickly, then it stays up.

Many platforms do not even have a policy on whether a post that is sponsored should be identified as such. The policies' silence allows advertisements to take place without any mention that they are, in fact, advertisements.[130] There are countless examples of users recommending a particular product, or posting a YouTube video about their experience with a particular product, without any mention that they are receiving payments or kickbacks to advertise that product.[131] While presented as spontaneous, independent, and even heartfelt recommendations, the individual was asked to promote this product in exchange for a financial benefit.[132] These posts, even if truthful in their content, are advertisements.

Moreover, the platforms' more recent attempts to root out deepfakes further normalizes the presence of deceptive content. Consider Meta's statement declaring its newly stepped-up efforts to eliminate deepfakes. The head of Meta's public policy team draws a distinction between manipulation that is benign and manipulation that intentionally misleads: "Some of that content is manipulated, often for benign reasons, like making a video sharper or audio clearer. But there are people who engage in media manipulation in order to mislead."[133] Intent aside, however, it is all manipulation. And, of course, much of what the platform calls "benign"—like the use of filters and editing, tools the platform encourages users to use—is *explicitly* designed to mislead, regardless of specific user preference.[134]

The policy solutions being developed to address deepfakes depend on preserving the distinction between types of deception. For example, the technology firm Adobe is working on a project, the "Content Authenticity

130. This problem pervades the wellness industry, the success of which has been buoyed by online platforms. *See* discussion *infra* Section V.C.1. Goop, a company founded by Gwyneth Paltrow, broke off its relationship with Condé Nast given the latter's requirement to separate ad content from informative content. *See* Taffy Brodesser-Akner, *How Goop's Haters Made Gwyneth Paltrow's Company Worth $250 Million*, N.Y. TIMES MAGAZINE (July 25, 2018), https://bit.ly/46psZ70 (explaining that "they weren't allowed to use the magazine as part of their 'contextual commerce' strategy" even though Goop "wanted to be able to sell Goop products (in addition to other products, just as they do on their site" and treat the "magazine customer [as] also a regular customer" of Goop).

131. *See* Roberts, *supra* note 21, at 120.

132. Of course, an influencer might be invited to promote a product they already used and loved; but failing to disclose the payment for promoting a product is still deceptive.

133. Bickert, *supra* note 48.

134. The academic conversation around deepfakes similarly asserts that shallow fakes are harmless and widespread. For example, Chesney and Citron define the problem posed by deepfakes by distinguishing them from the general deception that takes place online all the time: "[i]nnocuous doctoring of images—such as tweaks to lighting or the application of a filter to improve image quality—is ubiquitous." Chesney & Citron, *supra* note 2, at 1759.

Initiative," to allow users to see at a technical level whether a piece of content has been digitally manipulated.[135] While this might seem like an unqualified good, the tool risks giving viewers false confidence that the remaining material they are viewing has not been manipulated *in any way*, only because it has not been *digitally* manipulated. Yet more analog tools, like mislabeling a video, cropping the frame, and taking a short video clip out of context, can cause just as much harm as a digitally manipulated clip. In this way, rooting out deepfakes reifies the idea of an "authentic" media when, in fact, all media is subject to editing, interpretation, and contestation.

Ultimately, these policies all take place in the context of the platforms *authorizing* deception. At a basic level, these platforms allow, and encourage, users to alter images they post. That initial invitation—offered by the same companies that have promulgated policies against misleading content—is baked into the fabric of the platform, untouched by any of the policies that purport to root out deception.

The final picture that emerges is the following confused jumble of rules: misleading content is banned and discouraged, but only sometimes and only in some contexts; different rules apply if money is involved; and this is all subject to the caveat that it becomes a problem *only* if someone notices—and even if someone notices, the platform might not do anything about the deception and, ultimately, allow it. The result of this morass of policies is an information landscape that is deeply confusing and rife with fakery. In effect, the policies aimed at eradicating deception are problematic in a more fundamental way: they create the illusion that by accounting for and addressing the deception that takes place online, what remains is authentic and, for the most part, real. In this digital world, it can be nearly impossible to tell truth from deception.

IV. THE PROBLEM WITH SHALLOW FAKES

We have thus far described the prevalence of shallow fakes and identified how platforms create a market for engaging in online fakery. Now we turn normative. Should we be worried about shallow fakes? What is so problematic about self-enhancement or deception about the self?

Deception is fundamental to human interaction. We all lie a little. Research shows that we lie every day in our social interactions, with some estimates putting it at two lies per day.[136] More fundamentally, it is

135. *See* Eric Abent, *Adobe expands Content Authenticity Initiative tools to fight misinformation*, SLASHGEAR (Oct. 26, 2021, 8:12 AM), https://bit.ly/46eL7jE; *see also You decide what content to trust*, VERIFY, https://bit.ly/3r0ouzm (last visited Sept. 12, 2023).

136. *See* Bella M. DePaulo et al., *Lying in Everyday Life*, 70 J. PERSONALITY AND SOC. PSYCH. 979, 991 (1996) ("Participants in the community study, on the average, told a lie every day; participants in the college student study told two.").

understandable, even expected, that we would want to present the best versions of ourselves to our peers.[137] And, there might be real value in protecting the decision to assume a virtual identity as an important aspect of self-discovery or self-control.[138]

We do not dispute the existence, or at times even the necessity, of certain kinds of deception. Instead, we argue that the ecosystem in which people have incentives to fake many aspects of their digital selves has downsides worth taking seriously. In this Part, we outline three categories of harm. First, shallow fakes are deeply gendered, encouraging and propagating traditional gender roles and normative body types and exacerbating body issues, especially for young women and those who are gender-nonconforming. Second, many of the filters that propagate today's digital deception enable both racial appropriation and exclusion. The fact that platforms push users toward conformity is exactly the opposite of what we would expect to see in a medium that is purportedly designed for self-expression, like social media. Third, and finally, we explore some of the broader societal harms from living in a digital ecosystem marked by the kind of casual deception that takes place every day online.

One might think that if shallow fakes were as problematic as we argue, they never would have become so commonplace. But this logic is inverted. It is precisely *because* shallow fakes are so commonplace that they are problematic. The harms we describe are harms produced by the ubiquity of shallow fakes. We do not take issue with any single instance of shallow fakery but instead call attention to the aggregate effect of living in a world awash in shallow fakes.

Although we are interested in an aggregated problem, that does not mean that the harms are only felt in the aggregate. To the contrary, living in the world of shallow fakes can and does cause acute harms to individuals. This very dynamic—that the harm is felt at the individual level but is not inflicted at the individual level—is what makes shallow fakery difficult to remedy. One cannot pinpoint a single act of deception as the cause of someone's body dysmorphia or feelings of exclusion or any other consequent harm. Not only is this dynamic an obstacle to addressing

137. *See* GOFFMAN, *supra* note 45, at 35 (discussing "the tendency for performers to offer their observers an impression that is idealized in several different ways").

138. *See, e.g.*, Naramol (Jaja) Pipoppinyo, *Queer Identity Online: The Importance of TikTok and Other Media Platforms*, MEDIUM (Dec. 1, 2020), https://bit.ly/42XhjoW (describing "Gay TikTok" and its repudiation of "the mainstream, allowing [users] to embrace what it truly means to be queer"). *But see* Daniel Kershaw, *LGBTQ and Online Identities*, MEDIUM (Dec. 19, 2013), https://bit.ly/3NMan9N (noting the ways that gender play arises in online communities but also finding that "the community sometimes finds this form of gender [play] deceptive" and having a "fluid identity" can become overly idealistic).

the problem, but it is also a reason for why shallow fakes are so persistent and problematic.

A. Gendered Harms

There is a distinctly gendered nature to the harms produced by everyday deception online. When Frances Haugen, the Facebook whistleblower, testified before Congress in October of 2021, she revealed that the firm doggedly pursued teenage users, particularly girls, despite the fact that the firm's own research suggested a range of harms for those users.[139] As the Wall Street Journal explained in describing the firm's internal research, the firm repeatedly found "that Instagram is harmful for a sizable percentage of [young users], most notably teenage girls."[140]

But the range of gendered harms are broader and go much deeper. Shallow fakes are harmful to all marginalized genders. The political economy of seeking clicks, engagement, and an online audience reinforces traditional gender norms and excludes those who do not fit into the narrow binary presented. This is a complicated critique to undertake given that female influencers have gained particular traction in these online spaces, often by capitalizing on traditionally feminine roles, and have succeeded in monetizing activities that are otherwise excluded from the market.[141] The harms we identify are not the commercialization of previously "intimate" spaces. Rather, the harms follow from how these spaces get

139. *See Protecting Kids Online; Testimony From A Facebook Whistleblower: Hearing Before the Subcomm. on Consumer Protection, Product Safety, and Data Security of the S. Comm. on Commerce, Science, and Transportation*, 117 Cong. 9 (2021) (Statement of Frances Haugen) available at: https://bit.ly/3pelE9D [hereinafter *Testimony From A Facebook Whistleblower*]. Some have argued that even the attention paid to Frances Haugen as a whistleblower has a gendered and racist history. *See* Maria Farrell, *Two Facebook whistleblowers leaned in, but only one became a media star*, THE CONVERSATIONALIST, https://bit.ly/3r3Ycwb (last visited June 25, 2023) (arguing that only one whistleblower "was given the role of princess, commanding attention and praise," but in so doing reinforcing the accepted hierarchy of the importance of politics as compared to the relative unimportance of appearance, noting "Instagram's possible effects on their [Americans who work for the Wall Street Journal] adolescent daughters' self-image"). Before Frances Haugen there was Sophie Zhang, the daughter of Chinese immigrants. In deciding to become a whistleblower, she made the decision to hide her trans identity. Her 8,000 word memo focused more on Facebook's lack of attention to the political manipulation taking place online. *See* Karen Hao, *She risked everything to expose Facebook. Now she's telling her story.*, MIT TECH. REV. (July 29, 2021), https://bit.ly/44f58EY.

140. *See* Wells et al., *supra* note 21.

141. *See, e.g.*, Noah D. Zatz, *Working at the Boundaries of Markets: Prison Labor and the Economic Dimension of Employment Relationships*, 61 VAND. L. REV. 857, 916 (2008) ("Feminist scholarship has pioneered broader accounts of where economic activity, and particularly productive work, occurs. This research focuses principally on activities associated with the family—sexuality, reproduction, housework, and caregiving.").

constructed—in ways that reinforce stereotypically male and female roles and exclude those who do not fit into traditional gender scripts.

We begin this section by identifying some acute and immediate harms. We end by stepping back and taking stock of the different ways the platforms entrench normative performances of gender.

1. Body Dysmorphia

People who spend time on social media sites regularly have higher rates of body dysmorphia, which is defined as "a mental health condition in which you can't stop thinking about one or more perceived defects or flaws in your appearance."[142] As one researcher explained, "Smartphones, together with the cosmetics industry, are producing significant shifts in young women's visual literacies of the body—particularly the face—such that they quite literally see themselves and others differently from previous generations."[143] Facebook's research notes that this is especially true for the kinds of filtered images that are shared on Instagram, where it found that "[s]haring or viewing filtered selfies in stories made people feel worse."[144] These findings are not a niche concern, as Facebook reveals that it makes "body image issues worse for one in three teen girls."[145]

Contributing to the problem is simply the physical reality of taking a selfie. That is, "the angle and close distance at which selfies are taken may distort facial features and lead to dissatisfaction."[146] But it is made worse by online tools that allow people to manipulate their image with endless tweaks and filters. For example, one study found that adolescent girls who spent more time manipulating their photos reported higher levels of body dysmorphia than those who spent less time doing so.[147] This is merely a correlation; it is possible that the causal story is that girls with higher rates of body dysmorphia are likely to spend more time filtering their photos.[148]

142. MAYO CLINIC, *Body dysmorphic disorder*, https://bit.ly/3N4s83T (last visited Sept. 12, 2022).

143. Gill, *supra* note 26, at 35.

144. Wells et al., *supra* note 21.

145. *Id.*

146. Susruthi Rajanala et al., *Selfies-Living In The Era of Filtered Photographs*, 20 JAMA FACIAL PLASTIC SURGERY 443 (2018); *see also* Brittany Ward et al., *Nasal Distortion in Short-Distance Photographs: The Selfie Effect*, JAMA FACIAL PLASTIC SURGERY (Mar. 1, 2018), https://bit.ly/3rHHp2p.

147. *See* Sian A McLean et al., *Photoshopping the Selfie: Self Photo Editing and Photo Investment are Associated with Body Dissatisfaction in Adolescent Girls*, 48 INT'L J. EATING DISORDERS 1132 (Dec. 2015).

148. Girls generally tend to be more invested in their social media use than boys. *See* Jacqueline Nesi et al., *Emotional Responses to Social Media Experiences Among Adolescents: Longitudinal Associations with Depressive Symptoms*, J. CLIN. CHILD & ADOLESCENT PSYCH. 12 (2021) ("Girls reported more frequent positive and negative emotional responses to social media compared to boys."). Of course, online spaces may also provide a sense of community for these same teens, see Victoria Rideout et al., *Digital*

That would still be problematic, though, because the filtering tools seem to create a feedback loop for a slice of the population that is already susceptible to body dysmorphia.[149]

There is enough additional evidence to create a credible claim that the very existence of the filtering applications contributes to the incidence of dissatisfaction with one's body. Another study found that 32% of teenage girls said that when they felt bad about their bodies, Instagram made them feel worse.[150] In the words of one study participant, the feeling of seeing highly edited images of other women on Instagram is "like stick thin women with the most amazing butt and the most amazing long hair, and I'm just like, this isn't me, and why am I constantly seeing this? And it does make you feel abnormal, sometimes, and you are normal."[151] Another said, "[W]hen it comes to my skin, I know in my head that is normal. But when you see the content, it's like, it does make you feel almost abnormal because it's showing you that it shouldn't be that way."[152] The prevalence of shallow fakery redefines what is considered normal.

2. Depression and Anxiety

Young people, especially young girls, are under extraordinary pressure to conform to what they see on social media. The result is an epidemic of mental health problems. One Facebook study shows that 13.5% of teen girls reported suicidal thoughts increased since joining Instagram.[153] As one internal report conducted by Facebook noted, "[t]eens blame Instagram for increases in the rate of anxiety and depression."[154] This is, of course, related to body image and norm-conforming pressures described above. As one teen who was interviewed commented, "I see all these perfect bodies in bikinis and it makes me feel really low."[155]

3. Pressure to Sexualize

There is intense pressure, even for very young users, to present themselves as sexual beings. A majority of women surveyed about social media representations mentioned sexualization without being prompted.[156]

Health Practices, Social Media Use, and Mental Well-Being Among Teens and Young Adults in the U.S., PROVIDENCE ST. JOSEPH HEALTH DIGIT. COMMONS 15 (2018), but this does not minimize the harms caused by the shallow fakes found online.

149. *See generally* Gill, *supra* note 26.
150. *See* Wells et al., *supra* note 21.
151. Gill, *supra* note 26, at 19.
152. *Id.*
153. *See* Wells et al., *supra* note 21.
154. *Id.*
155. Gill, *supra* note 26, at 19.
156. *See id.*, at 42.

Online photographs are being edited not just to beautify but to sexualize. It is, in the words of one study, "a visually centered social media that involves the presence of sexualized imagery," which, in turn, has a negative impact on mental health.[157] Seventy-five percent of all young women in a recent survey said that they feel pressure to receive "likes" on their social media posts.[158] Another study found that "sexualized photos garnered more likes on Instagram," suggesting that teens feel pressure to post sexualized versions of themselves.[159] This leads to lower feelings of self-esteem and body image.[160] Such findings are widespread and are consistent with longstanding concerns about sexualization in traditional media and self-esteem problems, especially in younger people.[161]

To be clear, the harm we identify is the pressure to conform to the images presented online, a pressure which functions as a one-way ratchet. Some individuals might feel empowered to present sexualized images of themselves and therefore not experience it as a "harm" at all.[162] Our concern is not with these users but with the baseline the platforms create. All users are pushed to participate in this dynamic, which makes it less of a choice and more like the price of admission.[163] These platforms

157. Francesca Guizzo et al., *Instagram Sexualization: When Post Make You Feel Dissatisfied and Wanting to Change Your Body*, 39 BODY IMAGE 62, 62 (Dec. 2021).

158. Gill*, supra* note 26, at 24.

159. Laura Ramsey & Amber L. Horan, *Picture This: Women's Self-Sexualization in Photos on Social Media*, 133 PERSONALITY AND INDIVIDUAL DIFFERENCES 85, 85 (2018); *see also* Kun Yan et al., *A Sexy Post a Day Brings the "Likes" Your Way: A Content Analytic Investigation of Sexualization in Fraternity Instagram Posts*, 26 SEXUALITY & CULTURE 685, 685 (2022) (finding a positive association between the degree of sexualization in a post and the traffic and likes it received).

160. *See id.*

161. *See* Marika Skowronski et al., *Predicting Adolescents' Self-Objectification from Sexualized Video Game and Instagram Use: A Longitudinal Study*, 84 SEX ROLES 584, 585 (2021) (reporting the results of a longitudinal study involving 660 German adolescents and concluding that "sexualization in video games and on Instagram can play an important role in increasing body image concerns among adolescents"); *see also* Thomas Plieger et al., *The Association Between Sexism, Self-Sexualization, and the Evaluation of Sexy Photos on Instagram*, 12 FRONTIERS IN PSYCH. 1, 1 (Aug. 2021) (reporting results of a survey of 916 participants and finding that "there were substantial correlations between appropriateness and attractiveness evaluations of the presented photos and the self-sexualizing posting behavior and enjoyment of sexualization of female users").

162. *See, e.g.*, Emily Ratajkowski, *Emily Ratajkowski Explores What It Means to Be Hyper Feminine*, HARPER'S BAZAAR (Aug. 8, 2019), https://bit.ly/3WEQMLw. Ratajkowski explains that:

> Despite the countless experiences I've had in which I was made to feel extremely ashamed and, at times, even gross for playing with sexiness, it felt good to play with my feminine side then, and it still does now. I like feeling sexy in the way that makes me, personally, feel sexy. Period.

Id.

163. One response to this harm, as to all the harms we identify, is to remove oneself from social media entirely. While that might be a solution for some individuals, that is not directly responsive to the harms caused by social media platforms—those harms are what

perpetuate a distinctly modern feminist double-bind: women can become powerful players in a digital ecosystem, but the way they can achieve success is the timeworn one of relying on, and revealing, their body.[164]

Images that are sexualized are, of course, distinct from images that involve sex, which we do not identify as a harm.[165] Quite the opposite, in fact—we hope people find whatever satisfaction it is they are seeking online, sexual or otherwise.[166]

4. "Real-Life Filtered Look"

Cosmetic surgery rates have dramatically increased alongside the rise of social media, and many plastic surgeons now specialize in "Instagram Face."[167] Customers specifically seek to replicate how they appear on social media; they want a "real-life filtered look."[168] A 2021 survey by the American Academy of Facial Plastic and Reconstructive Surgery found that 77% of facial surgeons had treated customers in the previous year whose decision to undergo surgery was motivated by a desire to look better

---

our piece seeks to expose, with the aim of beginning a conversation on how to remedy them.

164. *See, e.g.*, RATAJKOWSKI, *supra* note 78, at 5. Ratajkowski writes:

> In many ways, I have been undeniably rewarded by capitalizing on my sexuality. I became internationally recognizable, amassed an audience of millions, and have made more money through endorsements and fashion campaigns than my parents (an English professor and a painting teacher) ever dreamed of earning in their lifetimes. I built a platform by sharing images of myself and my body online, making my body and subsequently my name recognizable, which, at least in part, gave me the ability to publish this book. But in other, less overt ways, I've felt objectified and limited by my position in the world as a so-called sex symbol . . . Whatever influence and status I've gained were only granted to me because I appealed to men.

Id.

165. Others, however, do. For a recent review of different feminist takes—negative and positive—on porn and its meteoric rise online, see SRINIVASAN, *supra* note 118, at 33–71.

166. *See, e.g.*, Emily Witt, *A Hookup App for the Emotionally Mature*, THE NEW YORKER (July 11, 2022), https://bit.ly/3C2Y8Pt (describing the author's experience on Feeld, a dating app that "is popular with nonbinary and trans people, married couples trying to spice up their sex lives, hard-core B.D.S.M. enthusiasts, and 'digisexuals,' who prefer their erotic contact with others mediated by a screen").

167. *See* Tolentino, *supra* note 58. Tolentino describes the "Instagram Face":

> It's a young face, of course, with poreless skin and plump, high cheekbones. It has catlike eyes and long, cartoonish lashes; it has a small, neat nose and full, lush lips . . . . The face is distinctly white but ambiguously ethnic—it suggests a *National Geographic* composite illustrating what Americans will look like in 2050 . . . .

Id.

168. American Academy of Facial Plastic Surgery, *AAFPRS Announces Annual Survey Results: Demand for Facial Plastic Surgery Skyrockets As Pandemic Drags On,* PR NEWSWIRE (Feb. 10, 2022), https://bit.ly/4518R9L [hereinafter *AAFPRS Survey*].

in selfies.[169] The tools offered by social media platforms present unrealistic images of bodies, often unattainable without a filter or surgery.[170] Yet, as beauty filters make people feel bad about their bodies, they also provide a template for how to fix them. These tools are having a profound effect on the cosmetic surgery market.

The autonomy and agency that are often lauded in discussions of plastic surgery[171] are less compelling when the ideals of beauty are being set by social media companies in uniform ways.[172] The debate over whether cosmetic surgery is ever the product of true agency or unadulterated choice runs deep.[173] Regardless of which side one falls on, evidence shows that cosmetic surgery is generally not undertaken to be beautiful, but to fit in—women engage in it as a way "to become ordinary, normal."[174] What is considered the norm, then, matters. And norm-setting is not value-neutral: "[n]ot every body will do; nor are all differences the same in Western culture."[175] Today's digital platforms are creating and

169. *See id.*

170. Low self-esteem is a major predictor in whether a woman will undergo plastic surgery. *See* Bonell et al., *The Cosmetic Surgery Paradox: Toward a Contemporary Understanding of Cosmetic Surgery Popularisation and Attitudes*, 38 BODY IMAGE 230, 233 (2021). Studies show, however, that "[d]espite the fact that low self-esteem directly predicts women's likelihood to undergo cosmetic surgery, research has indicated that longitudinal self-esteem improvements post-surgery are either small or non-significant." *Id.* at 234 (citations omitted).

171. *See* Melissa Febos, *The Feminist Case for Breast Reduction*, N.Y. TIMES MAGAZINE (May 10, 2022), https://bit.ly/3WDU5CJ (using breast reduction surgery as a stand-in for cosmetic surgery writ large, proposing that electing to undergo surgery is a feminist act, allowing her to "no longer [feel] so defined by [her] corporeal form").

172. The "Snapchat selfie" is something that plastic surgeons now specialize in. *See* Anna Davies, *People are getting surgery to look like their Snapchat selfies*, BBC THREE (Apr. 19, 2018), https://bit.ly/43tkYLM.

173. *See* Kathy Davis, *Revisiting Feminist Debates on Cosmetic Surgery: Some Reflections on Suffering, Agency, and Embodied Difference*, COSMETIC SURGERY: A FEMINIST PRIMER 38–39 (Cressida Heyes & Meredith Jones eds., 2016) (describing criticism of her work by feminist philosopher Susan Bordo, in particular the reliance on concepts of "agency" and "choice" and "freedom" in describing the women who undertake elective cosmetic surgery).

174. *Id.* at 36. The very existence of "[c]osmetic surgery is predicated upon definitions of physical normality," and these "categories of 'normality' and 'abnormality' are drawn upon in both medical discourse on cosmetic surgery . . . and in individuals' accounts of their surgical experiences." *Id.* at 43.

175. *Id.* at 45. Davis criticizes the discourse that presents "cosmetic surgery . . . as neutral technology, ideally suited in altering the body in accordance with an individual's personal preferences" because, among other problems, "[i]t discounts the universality of white, Western norms of appearance, which shape individuals' perceptions of what they consider to be desirable appearance as well as the kinds of interventions that are deemed acceptable." *Id.* at 44–45. We are mostly discussing these questions from an American point of view, and thus Western culture matters. Beauty ideals, however, of course differ across the world and it is often the case that "*local* cultural dynamics trump[] outside influences when it comes to health and beauty norms." HU, *supra* note 90, at 152. Elise Hu has considered the popularity of plastic surgery in South Korea specifically and has

perpetuating these norms en masse, as evidenced by the ubiquity of the requests for a particular kind of filtered look.

The norm-creation these platforms are engaged in extends to the procedure of cosmetic surgery itself. TikTok is replete with videos demonstrating before and after pictures of procedures like nose jobs.[176] The black eyes, broken nose, and nose bleeds that follow these procedures are included in 15-second clips that culminate with a celebratory "'nose job check' sound."[177] Currently, the TikTok hashtag #nosejob has over 1.6 billion views.[178]

The problem we identify is not the desire for cosmetic surgery, nor even the cosmetic surgery itself. We also roundly reject the notion that these desires or discussions are "frivolous"[179]—indeed, our whole concern is with a swath of behavior that could be, and is, easily dismissed as such. Instead, our critique is of the conditions that have led to a pronounced increase in these elective procedures along with the specific nature of the requests,[180] namely, to mimic a highly stylized image offered by the digital platforms that is unattainable without filters or, in the physical world,

---

explained how definitions of beauty are informed internally, making it "colonialist to emphasize race as a determining factor in the beauty decisions of Asians." *Id.* at 152–53. She does note that current ideals of being "cute" in South Korea are informed by "anime characters or the enhancements from digital filters on . . . social apps." *Id.* at 158.

176. On TikTok, "[p]lastic surgery videos are endemic . . . . The hashtag #plasticsurgery has over 3.8 billion views" while "[t]he hashtag #nosejobcheck, which mainly consists of videos showcasing before-and-after clips of nasal surgery, has accumulated over one billion views on the platform." Joshua Zitser, *Insider created a tiktok account and set the age at 14 to test how long before a plastic surgeon's promotional video appeared. It only took 8 minutes*, INSIDER (Jan. 10, 2021, 10:41 AM), https://bit.ly/3C5lLa2.

177. These videos are visible to young users, and TikTok's community guidelines do not prevent them from being shown as long as they are "not 'graphic' and don't contain 'gore.'" *Id.* Charli D'Amelio, one of TikTok's most viewed users, documented her nose job when she was 16. The caption reads: "two broken noses lots of nose bleeds and breathing problems for 11 months! I can finally breathe like normal and get back to dancing." She then tags her surgeon, Dr. Kanodia, @drkanodia90210. Charli D'Amelio (@charlidamelio), TIKTOK (Jul. 28, 2020), https://bit.ly/3N4cdT1.

178. *See* Zitser, *supra* note 176.

179. Davis, *supra* note 173, at 87 (describing Iris Young's criticism that "much of the cosmetic surgery women undergo must be 'frivolous and unnecessary, like diamonds or furs'") (internal citation omitted).

180. The desire to change one's body was exacerbated by the COVID-19 pandemic because people were spending time online using applications that did not have filters, which is believed to have caused a considerable rise in elective cosmetic surgeries. *AAFPRS Survey*, *supra* note 168 ("Unlike selfies and video editing apps like TikTok and Reels on Instagram, the video conferencing used for school, work and ZOOMing with family and friends does not allow for filtering capabilities, making it a particularly easy lens for self-scrutiny."). As the President of the leading professional association for facial plastic surgeons noted, "[r]eal time video cannot be FaceTuned or photoshopped to smooth out a bump on the nose, crow's feet or a sagging neck." *Id.*

cosmetic surgery.[181] The harm we are surfacing is of social media platform algorithms driving a very particular, very limited standard of beauty, which is causing people to see themselves differently[182] and to change their bodies and their faces as a result.[183]

5. Reinforcing Traditional Gender Roles

Online fakery in general reinforces rather than disrupts traditional gender norms. A survey of teenage girls revealed "a very high degree of consensus about how women 'should' look: 'no body hair, white teeth, curvy but slim, good skin.'"[184] The same report found that young women feel pressure to have a certain look and that such pressures are especially hard for anyone who is nonbinary. Nearly everyone is chasing a look that is "too perfect, too sexualised, too white, too heteronormative, too middle class and do[es] not represent the lives of disabled and/or nonbinary people."[185] Indeed, not just the selves but the lives that users present online are stereotypically heteronormative. One lesbian survey respondent noted that "[w]omen are so often presented in relation to men (as a wife, girlfriend, sex object or in a heterosexual family), [that] as a lesbian woman I don't relate to these images."[186]

Much of the fakery that people engage in online is something people do to themselves.[187] But the filter compounds this process by changing appearances in ways that reinforce pre-existing gender norms. For example, the beautification filters that have become increasingly common

181. Lele Pons, who is ranked as the number-one, highest-paid non-celebrity on Instagram, is very open about her plastic surgery and posts before and after pictures of herself on her site. *See* Lele Pons (@lelepons), INSTAGRAM (March 5, 2022), https://bit.ly/3INe38v.

182. *See* Tate Ryan-Mosley, *Beauty filters are changing the way young girls see themselves*, MIT TECH. REV. (Apr. 2, 2021), https://bit.ly/45ARxcD.

183. *See* HU, *supra* note 90, at 135 ("The technological gaze keeps feeding us an ever-evolving, ever-narrowing beauty ideal. With enough money, we use ever-improving artificial implants, injections, or surgery to look however we want, fueling a filter-to-filler pipeline.").

184. Gill, *supra* note 26, at 26.

185. *Id.* at 8. Another person described the conformity as follows: "A very fit, slim body has been normalised as what a woman should look like. Men also face the normalisation of a fit physique." *Id.* at 29; *see also Por que queremos ser iguales?*, LA VANGUARDIA (June 29, 2015), https://bit.ly/3qZhkeS (discussing the criticisms of the Miss Spain 2022 pageant for picking three finalists who all looked very similar and addressing the paradox of living in a society that is at once demanding more equality and greater diversity but also experiencing a push towards a singular definition of beauty represented on social media networks).

186. Gill, *supra* note 26, at 45.

187. Even if they describe disliking it. As one influencer notes on her Instagram stories: "I literally just always use a Paris filter. Like it's just habit at this point." In the next slide, she writes "Here with Paris filter. These filters fuck you up." *See* Sara Foster (@sarafoster), INSTAGRAM (July 6, 2022).

on apps are designed to identify a female face and make it more slender, while they make a face identified as male more broad.[188] These same filters try to make women's bodies thinner and men's more muscular.

There are counter examples, to be sure. A recent report described how Instagram can be a "lifeline" for nonbinary people "struggling to find others just like them."[189] The same report noted, however, that the space is filled with risks for nonbinary people, who post pictures of themselves only to be criticized and bullied for not conforming to gender stereotypes.[190] One five-year study of queer youth of color found that social media platforms like Facebook were dangerously heteronormative and created spaces of "default publicness," resulting in offline harms like being disowned from one's family.[191]

The marketplace for clicks and sponsored content pushes individuals into more traditional roles if they want to succeed. Content online exists not only in a market for likes but in an actual market where platforms are seeking payment, and individual users are seeking endorsements. Many brands use algorithms to help identify which individuals they should contact to sponsor their products online. Studies of these algorithms reveal how certain terms and activities are categorized in ways that marginalize already-marginalized users. For example, Peg, a UK-based tool that enables brands to identify possible marketers, codes the use of the term "queer" as profanity, making individuals who employ that term less attractive to brand partnerships and thereby excluding them from the market and from the ability to influence.[192] Similarly, YouTube creators who identify as LGBTQ+ have a difficult time monetizing their content given that such content is often age-restricted and marked as "not being 'advertiser and family friendly.'"[193]

---

188. *See* Sage Anderson, *Snapchat's 'gender-swap' filter exposes the internet's casual transphobia*, MASHABLE (May 16, 2019), http://bitly.ws/F59W.

189. Wortham, *supra* note 30.

190. *See id.* (interviewing one user who "doesn't identify as any gender" who spoke to the possibilities that social media creates of rendering gender nonconforming lives visible and of challenging "mainstream perceptions of gender," while also noting the violence such individuals face, explaining, "I still receive daily hate mail from people of all genders telling me that my body hair is ugly & that I need to shave to be more 'real' & 'beautiful'") (internal quotation marks omitted).

191. *See* Alexander Cho, *Default Publicness: Queer Youth of Color, Social Media, and Being Outed By The Machine*, 1 NEW MEDIA & SOC. 3183, 3184, 3187 (2017) (contrasting Facebook to Tumblr, and showing that queer young people preferred Tumblr, which has less of a default public character).

192. *See* Sophie Bishop, *Influencer Management Tools: Algorithmic Cultures, Brand Safety, and Bias*, 7 SOCIAL MEDIA & SOCIETY 1, 8 (March 30, 2021), http://bitly.ws/F5ar (explaining "[w]hile queer does have roots as a homophobic slur, it is a term widely used in activism and LGBTQ+ communities, in addition to within deconstructivist theory to recognize that sexualities are 'unstable, fluid, and constructed") (citation omitted).

193. Zoë Glatt & Sarah Banet-Weiser, *Productive Ambivalence, Economies of Visibility and the Political Potential of Feminist YouTubers*, ASSOCIATION OF INTERNET

Even acts that appear to subvert heteronormative ideals end up propping them up. Take the example of women who present themselves as mothers online. "Mommy bloggers" have been largely replaced by mommy Instagrammers, who, instead of sharing "the ups and downs of their lives," now post about "trips to Greece and carefully arranged living rooms."[194] Instagram is, by its very nature, more image-based than narrative-based and feeds into a particular ecosystem that relies on "likes and hearts."[195] Minor insurrections against the demands of motherhood appear as hashtag #momfails or as purported displays of a less-than-perfect life. Most of these "failures," however, involve leaning into the role of perfect mother and continue to portray largely unattainable goals. Research into situations in which moms cast themselves as a "hot mess" or describe their homes as chaotic shows that they are neither: a "visual inspection of these posts revealed clean and well organized homes[] and women who were well kempt (e.g., styled and coloured hair, wearing cosmetics, wearing clean, nice clothes, accessorized outfits with jewellery, etc.)."[196]

These admissions of imperfection are often paired with product placement. Consider, for instance, a picture of a young mother's ten-day postpartum belly, seemingly awash in one of Instagram's filters, accompanied by the caption "[s]o proud of my body and the life it has created."[197] The image is at once intended to be relatable—the belly is still visible after giving birth—but also aspirational in that there are no other signs of the baby on the body, which is tan and blonde, positioned against a bathroom that is spotless and framed as place for relaxation (candles and a loofa are visible in the background).[198] The caption further expresses gentle self-acceptance: "giving myself all the love, care, hot showers, and R&R it needs."[199] Nowhere does it state that the post is sponsored. Yet the

RESEARCHERS (AOIR) VIRTUAL CONFERENCE (2020); *see also* Ari Ezra Waldman, *Disorderly Content*, 97 WASH. L. REV. 907, 910–11 (2022) (arguing that content moderation maintains and reifies "social media as 'straight spaces' that are hostile to queer, nonnormative expression").

194. "The death of the mom blog has something to do with shifts in how people consume and create on the Internet. Blogging on the whole has fizzled as audiences and writers have moved to other platforms." Sarah Pulliam Bailey, *What ever happened to the mommy blog?*, CHICAGO TRIBUNE (Jan. 29, 2018), https://bit.ly/45duS5x. Those platforms include Instagram, which "is built for beauty (its filters make your life look better), not for rawness." *Id.*

195. *Id.* ("The shift to shorter posts and an emphasis on likes and hearts has changed the tone and content of what moms find online: more pictures, fewer words, less grit.").

196. Kelly D. Harding et al., *#sendwine: An Analysis of Motherhood, Alcohol Use and #winemom Culture on Instagram*, 15 SUBSTANCE ABUSE: RESEARCH AND TREATMENT 1, 4–5 (2021).

197. Pamela Tick (@pamelatick), INSTAGRAM (Dec. 17, 2021), http://bitly.ws/F5fN.

198. *See id.*

199. *Id.*

image includes a clickable link to Spanx, a company that "specializes in foundation garments intended to make people appear thinner."[200] The message of self-love is thus in service to a brand whose purpose is to minimize the very bulge the mom is outwardly celebrating, making her fit into the conventional—slim—body image that women, including those who just gave birth, are pressured to achieve.[201]

This is no coincidence: promoting self-care goes hand-in-hand with product placement.[202] Motherhood is so well-represented on social media platforms because women make decisions over household purchases and companies are eager to reach them as audience members and consumers.[203] In exploring the use of alcohol on Instagram by mothers, one study found that posts tend to position wine as a means of coping with overwhelming "domestic and motherly responsibilities."[204] Rather than question the conditions that lead the allocation of caretaking responsibilities to fall entirely on "mom," or address any potential underlying health or mental issues, the content presents a product, wine, as a way of "bringing serenity and peace for women."[205] Wine becomes a "part of routine self-care, and . . . a necessary way to cope with being a modern mother."[206]

While monetizing work within the home is in some ways preferable to the alternative of keeping "women's work" gratuitous and outside of a cognizable market exchange, the kind of work that is recognized is

200. *Spanx*, WIKIPEDIA, https://bit.ly/46jYBuC.

201. Breastfeeding challenges and feeding complications likewise appear in conjunction with a recommendation for a specific powdered milk product. *See* Daphne Thompson*, Hannah Bronfman Shares Her Infant Feeding Journey & What Led Her to Bobbie*, MAMAZINE (Aug. 18, 2021), https://bit.ly/3rystmW (discussing Bronfman's "feeding journey" in conjunction with @bobbie, a company "founded by moms"). *See also* Melissa Murray, *Foreword: The Milkmaid's Tale*, 57 CAL. W.L. REV. 210, 227–28 (2021) (describing the pressure imposed on mothers to conform and that "[i]n a culture in which 'breast is best'—and indeed, is so self-evidently correct that alternatives are never even broached or contemplated—it is easy (or at least easier) to condemn those who depart from the orthodoxy as deviant" especially "when these individuals are Black women").

202. When mothers engage in mini-rebellions and challenge the norms of how a "good mother" acts, their revolutions stall in the register of self-care. In the context of wine drinking on Instagram, the following findings have been made:

> #winemom culture has enabled momtrepreneurs, women who are balancing the roles of business owner and parent, to create an online persona that can be used on social media sites to provide one-stop-shops that sell products while allowing women to partake in an online community of like-minded individuals with the shared experience of motherhood. The creation of these online communities has allowed wine companies (amongst other companies) to market to women without seeming too obvious or obnoxious.

Harding et al., *supra* note 196, at 2.

203. *See id.*

204. *Id.* at 6.

205. *Id.* at 8.

206. *Id.*

necessarily marked as "woman's work" in these online spaces. And the individuals who are most successful in this specific market are middle-class, white women who continue to be most legible as mothers.[207] Not all mothers can engage in this market on the same terms, or at all. These posts participate in constructing an ideal mother and an ideal consumer who is "white, middle class, [and] cisgender."[208]

B. Racialized Harms

As the discussion of gender makes clear, users construct their images online in ways that implicate race. In this section, we identify further racial harms that include blackfishing, whitewashing, and other forms of appropriation and exclusion. To a certain extent, of course, this reflects the interests of users—but the platforms enable such practices by offering specific filters that amplify and propagate each of the harms we identify.

1. Blackfishing and Other Forms of Appropriation

Blackfishing is a form of cultural appropriation in which someone who is not Black acts or appears to be Black in order to acquire some sort of capital—to appear more attractive or garner attention or acquire financial gain.[209] The concern underlying blackfishing is that someone can tweak their image in ways that allow them to be Black, without actually being Black, and in this manner can pick and choose the benefits of being Black without experiencing any of the drawbacks. As one critic suggests, "[i]nstead of appreciating Black culture from the sidelines, there's this need to own it, to participate in it without wanting the full experience of Blackness and the systemic discrimination that comes with it."[210] While concerns over cultural appropriation have existed for centuries, blackfishing is a phenomenon distinctly tied to social media. Its prevalence is directly related to the rise of filters and other software that allow people to change their appearance by "using image editing tools to darken their complexion or change their facial features to appear more Black."[211] It is

207. *See id.* at 7 (discussing how "wine mom culture" on social media leads to "a continued reproduction of white, middle-upper class, neoliberal values").

208. *Id.* The study explains: "It is notable that only one image corresponded to a visible woman of colour, with all other images aligning with predominantly white women who present themselves as being affluent." *Id.*

209. *See* Stevens, *supra* note 31, at 1 (defining blackfishing as "a practice in which cultural and economic agents appropriate Black culture and urban aesthetics in an effort to capitalize on Black markets"). For a nuanced take on how race can be considered mutable, see Deepa Das Acevedo, *(Im)mutable Race?*, 116 NW. U. L. REV. ONLINE 88 (2021).

210. Faith Karimi, *What 'Blackfishing' means and why people do it*, CNN (July 8, 2021, 8:37 AM), https://bit.ly/3MlOfRI.

211. Zawn Villines, *What to know about blackfishing*, MED. NEWS TODAY (Nov. 9, 2021), https://bit.ly/3OA5d1p.

also apparently successful: three of the top ten Instagram earners are routinely accused of blackfishing.[212]

Examples abound. Emma Hallberg is a Swedish model and influencer. She has over half a million followers on Instagram and she is known for her makeup tutorials, which showcase her skin's bronzed glow.[213] Most people assumed that Emma was Black, but, it turns out, that she identifies as white. As she told BuzzFeed, "I do not see myself as anything else than white," and "I get a deep tan naturally from the sun."[214] Emma was accused, along with other white Instagram models, of "adopting what some have called digital blackface, altering their appearance with makeup and using Afrocentric hairstyles."[215] The purpose in doing so is "to build their personal brand and secure lucrative brand endorsements"—which, for many users, is *the* reason for being on social media.[216]

Appropriation can take place not only in terms of physical appearance but also in terms of activity. Addison Rae, a famous TikToker, has performed and popularized a series of dances authored by Black creators, without always giving them credit.[217] The scale of blackfishing on TikTok became significant enough that in 2021, Black TikTok creators staged a strike.[218]

Blackfishing is, in many ways, only the latest chapter in a very "old story about white people profiting off of black aesthetics to project a sense

212. The top-earners are Kim Kardashian, Ariana Grande, and Kylie Jenner. *See* Donna Tang, *How Much Do Instagram Influencers Make*, CREDITDONKEY, (Apr. 12, 2022), https://bit.ly/3q4NVPW; Stevens, *supra* note 209, at 1 (listing Kim Kardashian and Ariana Grande as being accused of blackfishing); Ryan Schocket, *Kylie Jenner is Being Accused of Blackfishing And the Twitter Reactions Say It All*, BUZZFEED (Oct. 23, 2021), https://bit.ly/45jYVsF.

213. *See* Tanya Chen, *A White Teen Is Denying She Is "Posing" As A Black Woman On Instagram After Followers Said They Felt Duped*, BUZZFEED NEWS (Nov. 13, 2018, 5:05 PM), https://bit.ly/3Wt26dN.

214. *Id.*

215. Stevens, *supra* note 31, at 1.

216. *Id.*

217. Addison Rae received backlash for performing dances on *Jimmy Fallon* and not giving credit to the choreographers. She does not claim otherwise and lists the creators on her YouTube channel. Describing them, she has said, "They're all so talented and I definitely don't do them justice." Joe Price, *Addison Rae Under Fire for Not Crediting Black TikTok Creators While Performing Challenges on 'Fallon' (Update)*, COMPLEX (Mar. 29, 2021), https://bit.ly/3pWoqjy.

218. *See* Sharon Pruitt-Young, *Black TikTok Creators are on Strike to Protest a Lack of Credit For Their Work*, NAT. PUB. RADIO (July 1, 2021, 11:00 PM), https://bit.ly/43cJQr8. TikTok claims it is taking steps to reduce the ability to engage in cultural appropriation, but Black users say there has been little change. *See* Vanessa Pappas & Kudzi Chikumbu, *A message to our Black community*, TIKTOK (June 1, 2020), https://bit.ly/45teQoA; Kalhan Rosenblatt, *Months after TikTok apologized to Black creators, many say little has changed*, NBC NEWS (Feb. 9, 2021, 5:11 AM), https://bit.ly/3Ixcmf6.

of edge without feeling any of the associated struggle."[219] But filters contribute to, and accelerate, this process by making these alterations easily available. The nature of being online also makes the changes difficult to identify when the only image one has access to is of a filtered individual. The market for likes, and the paid sponsorships that permeate the platforms, facilitate the use of Black culture "to promote the products they endorse in their posts and to increase their following."[220]

The platforms themselves also directly traffic in clearly stereotypical portrayals of different cultures and ethnicities. Instagram used to offer a filter called "Choco skin" which allowed users to automatically darken their skin and hair.[221] Other filters were aimed at making users appear more Asian, with names like "Asian Beauty" and "Geisha."[222] Snapchat has repeatedly offered filters that allow users to transform their faces into cartoon versions of Jamaican Rastafarians (the "Bob Marley" filter) and had one that was "anime-inspired," in which the user's face would be augmented with a rice hat, squinted eyes, and buckteeth.[223] Snapchat has since removed both filters.

Of course, invoking a culture that is not one's own is not per se problematic. And there are important expressive benefits to seeing different ethnicities represented on these digital platforms. But there are ways to pay homage to another culture without portraying it as one's own for financial gain or by reducing it to offensive stereotypes.[224] A platform offering users a Geisha filter that automatically allows them to digitally dress up as a Geisha is not that.

2. Whitewashing and Exclusion

While digital image enhancing tools allow people to appear darker than they are, filters also automatically whiten darker skin in photos. For example, the popular filtering app FaceApp sparked an outcry for its "Hot"

219. Spencer Kornhaber, *How Ariana Grande Fell Off the Cultural-Appropriation Tightrope*, ATLANTIC (Jan. 23, 2019), https://bit.ly/43kPG9R.

220. Stevens, *supra* note 31, at 6.

221. Charlie Duffield, *Instagram 'choco skin' filter is form of brownface, says teacher and activist*, EVENING STANDARD (Sept. 1, 2020), https://bit.ly/43gV5iR.

222. Sarah Lee, *Instagram Filters: 'Our skin is for life, not for likes'*, BBC NEWS (Oct. 19, 2020), https://bit.ly/3oG4Yav.

223. *See* Robinson Meyer, *The Repeated Racism of Snapchat*, ATLANTIC (Aug. 13, 2016), https://bit.ly/3WMsT4P; *see also* Sam Levin, *Snapchat faces backlash over filters that promotes racist stereotypes of Asians*, THE GUARDIAN (Aug. 10, 2016, 2:04 PM), https://bit.ly/43jmdxp.

224. Rihanna's decision to wear Chinese couture at the 2015 Met Gala is a good example; she used her platform to respectfully honor Chinese designs, and she gave credit to those designers. *See* Jenni Avins & Quartz, *The Dos and Don'ts of Cultural Appropriation*, THE ATLANTIC (Oct. 20, 2015), https://bit.ly/45F0kdE. The setting was also relevant: it was a costume ball at the Met Museum, and she chose a designer whose work was on display at the Museum. *See id.*

filter which made darker skin tones appear lighter.[225] The app's filter did not promise to make users whiter, but rather to make them "hotter," which, according to the platform, meant having whiter skin.[226] The same is true for Snapchat's "flower crown" filter, which one would expect to merely add a flower crown to images but which also whitens one's skin tone considerably.[227] Similarly, Instagram's "Attraction" filter—which has been used in over 143,000 videos—pushes people towards European standards of beauty.[228]

Whitewashing and blackfishing are not as paradoxical as might initially appear. The problem with blackfishing is that it allows individuals to selectively choose which aspects of Black culture to claim, in a world that still holds mainstream white beauty as the norm.[229] Indeed, the baseline, "unfiltered" look is to whiten and lighten. The cumulative effect of these filters is that they regularly exclude people of color.[230] For example, the "Glow" filter on TikTok, which is designed to make a face look more beautiful, simply does not work on some people of color.[231] As one TikTok creator put it, "my first reaction was like, 'Oh, great, another one of those beauty filters that changes our features to make us cater to the European so-called beauty standards."[232] The Glow filter has been used on over 3 million TikTok videos.[233] As one Myanmarese TikTok user explained, "You have to be a white woman. You have to have darker skin almost, but in the, bronze-y, 'white woman with a tan' way rather than like, actually working for people with different skin tones."[234] Even though the user base is diverse, the tools offered by the platforms are not.[235]

---

225. *See* Zoldan, *supra* note 32.

226. *See id.*

227. *See* Prakash, *supra* note 32.

228. *See* Jennimai Nguyen, *TikTok beauty filters can be super realistic-unless you're a person of color*, MASHABLE (Aug. 6, 2021), https://bit.ly/3ML5xrI.

229. *See* Stevens, *supra* note 31, at 12 ("The light-skinned, loosely textured hair and plump lips featured in their content are indicative of normative White beauty standards as a measure to which the black feminine body should aspire.").

230. *See* Nguyen, *supra* note 228.

231. *See id.*

232. *Id.*

233. *See id.*

234. *Id.*

235. *See id.* These effects are embodied in the kinds of cosmetic procedures that are pursued. As Kathy Davis has observed in her sustained study of cosmetic surgery:

> A critique of cosmetic surgery and, more generally, a politics of the body, cannot be reduced to *either* gender *or* race. An exclusive focus on gender would be inadequate for understanding why the practice of cosmetic surgery has been a primarily white, western enterprise. By the same token, an exclusive focus on race or ethnicity could not account for the fact that most operations on 'Jewish noses' or 'Oriental eyelids' are performed on women.

Kathy Davis, *Surgical Passing: Or Why Michael Jackson's Nose Makes "Us" Uneasy*, 4 FEM. THEORY 73, 85 (2003).

C. Democratic Harms

We have not meant to imply in our assessments of the harms that users are naïve and unsuspecting victims. In fact, because everything is being faked all the time, many of us view and engage in the online sphere with a certain cynicism. "That can't be real," or "they're just trying to sell something," are typical everyday reactions to a medium in which everyone is searching for clicks and scrolls. This reaction is heightened given that all of these interactions take place against a background of commercialization.[236] Knowing that "goods"—a product, a lifestyle, a better version of one's self—are constantly being peddled on social media means that people react skeptically to posts. The corollary worry about widespread deception is that no one trusts anyone about anything.

This skepticism manifests itself in a number of important ways that threaten a healthy, functioning democratic society. We identify two: the erosion of expertise and the inability to engage in informed public discourse. While others have raised these concerns in the context of fake news, their relationship with shallow fakes has not been considered. We do that here. In addressing the erosion of expertise, we consider the rise of the wellness industry—which has truly taken off on social media platforms—as a case study. In considering the impoverishment of our ability to engage in productive dialogue, we examine the role that constant, casual, and unchecked deception plays on the internet.

1. The Erosion of Expertise

One of the appeals of social media is its democratic nature: anyone can have a voice. Users know that any individual with a smart device and an internet connection can join in the discourse—there is no test to pass or requirement to satisfy prior to endorsing a perspective or stating an opinion.[237] It is precisely because everyone can speak that it is harder for authorities on a particular topic to stand out. The result has meant an erosion of public trust in experts. The Pew Research Center has shown that

236. *See* WU, *supra* note 10, at 5.

237. An exception to the lack of correctives on any information presented, and a turn towards expertise, happened during the COVID-19 pandemic. *See Helping People Stay Safe and Informed about COVID-19 Vaccines*, INSTAGRAM BLOG (Mar. 16, 2021), https://bit.ly/3qoZjWX (stating Instagram's practices of removing posts that violate its COVID-19 and vaccine policies and attaching COVID-19 and vaccine information labels to posts containing potentially unfounded health claims that direct users to "more credible information from health experts including the WHO and the CDC").

Americans' trust in experts is declining.[238] Other research has shown that this decline in trust is related specifically to social media use.[239]

The visual medium we have described is one where people present, and are presented with, unreal and unrealistic images. Routine exposure to such images leads many users to feel bad about themselves and to believe they can make small changes (edits, touch-ups, glow-ups) to their bodies or their lives to better themselves. In short, the world of shallow fakes makes self-improvement virtually a necessity. It is no surprise that the wellness industry has thrived on social media. Not only does the industry provide the tools for betterment of the self, but the platforms provide fertile ground for claims about various health and beauty products that go unchecked.

The core feature of the wellness movement is the permission to focus on, and obsess over, the presentation of the self. [240] Similar to how shallow fakes function, the wellness industry simultaneously manufactures the desire and supplies the tools for betterment. As writer Molly Young detailed in a deeply studied profile of a "moon juice" purveyor:

> What Goop (and acolytes like Moon Juice) sell is the notion that it's not only excusable but worthy for a person to spend hours a day focused on her tiniest mood shifts, food choices, beauty rituals, exercise habits, bathing routines and sleep schedule. What they sell is self-absorption as the ultimate luxury product.[241]

This turn inward means that the external world of facts and knowledge matters less.[242] The wellness industry depends on this turn—on the ability to disseminate opinions as knowledge and to sell goods under the guise of presenting facts. Take Goop, the lifestyle brand started by Gwyneth Paltrow: it began as a newsletter and has grown into a quarter-billion dollar business.[243] Goop is built on the concept of self-improvement

238. *See* Brian Kennedy et al., *Americans' Trust in Scientists, Other Groups Declines*, PEW RSCH. CTR. (Feb. 15, 2022), https://bit.ly/3C8W1tG.

239. *See* Dominik Andrzej Stecula et al., *How trust in experts and media use affect acceptance of common anti-vaccination claims*, MISINFORMATION REV. (Jan. 14, 2020), https://bit.ly/3MOTD05.

240. Shallow fakes are of a piece with the obsession over self-improvement and self-care. *See* JIA TOLENTINO, TRICK MIRROR: REFLECTIONS ON SELF-DELUSION 80 (2019) ("Old requirements, instead of being over-thrown, are rebranded. Beauty work is labeled as 'self-care' to make it sound progressive.").

241. Young, *supra* note 34.

242. *See* Peter Dahlgren, Commentary, *Public Sphere Participation Online: The Ambiguities of Affect*, 12 INT'L J. OF COMMC'N 2052, 2065 (2018) ("Today, in the viral world of online information . . . what we feel—is clearly on the rise. Truth becomes reconfigured as an inner subjective reality with an affective leap and thus becomes the foundation for validity claims about reality.").

243. *See* Brodesser-Akner, *supra* note 130.

through self-care.[244] It sells products like the "$66 'Jade Egg,'" a stone which is meant to be inserted into the vagina and which Goop claims "could balance hormones, regulate menstrual cycles, prevent uterine prolapse, and increase bladder control." None of these claims are supported by scientific evidence.[245] As a result of these and other assertions, ten California District Attorney's offices filed suit for false and misleading advertisement in violation of the state Business and Professions Code, section 17500.[246] Goop settled for $145,000.[247]

These deceptive marketing practices might be dismissed as the kind that happen in all industries. But the wellness industry is particularly rife with deceptive and unfair trade practices and is part of the rising tide of online misinformation.[248] As one report set forth, "[a] surge in misinformation has grown with the internet, making wellness strategies appear to have scientific foundations when instead they're fueling baseless and sometimes harmful theories."[249] To take another related example, vitamin sales, which are heavily promoted on social media, have jumped 40% from 2019 to 2020, despite little medical evidence to show that they increase health or wellness.[250]

Of course, traditional media also sells beauty, glamour, and the products that help achieve both. But there are journalistic standards they must adhere to that are wholly absent in the virtual arena. In 2017, Condé Nast, a global mass media company that owns Vogue, a fashion magazine marketed to women,[251] decided to partner with Goop to deliver content to Vogue's readers.[252] The deal soon fell apart. The problem was twofold. First, Vogue publishes a magazine, not a catalog, which means that it must enforce a separation between content and product placement—and therefore a separation between reader and consumer—which Goop does not do.[253] Second, Vogue requires fact-checking and support for scientific claims made.[254] Goop, however, understands that support to be

244. *See id.*

245. *See* Bill Bostock, *Gwyneth Paltrow's Goop settles $145,000 lawsuit over baseless vaginal eggs health claims*, BUS. INSIDER (Sep. 5, 2018, 7:15 AM), https://bit.ly/42isoR8 (noting that the settlement also related to an essential oil that Goop claimed would cure depression).

246. *See* Complaint at 3, People v. Goop, Inc. (Cal. App. 4th Supp. Aug. 31, 2018). No. 18CV001176.

247. *See id.*

248. *See* Marisa Fernandez, *Beware the "science" behind some wellness industry's claims*, AXIOS (Feb. 15, 2020), https://bit.ly/3OQzn0J.

249. *Id.*

250. *See id.*

251. *See Vogue (magazine)*, WIKIPEDIA, https://bit.ly/43z2rxA.

252. *See* Brodesser-Akner, *supra* note 130.

253. *See id.*

254. *See id.* As the profile on Goop and Gwyneth Paltrow ("G.P.") explains:

unnecessary, given that "they're never asserting anything like a *fact*" but rather "just asking unconventional sources some interesting questions."[255] Facts, which remain relevant to traditional media, are of diminishing importance to social media.

2. The Erosion of Public Discourse

The disregard for facts in the social media space encourages a distrust of the information presented in ways that affect our most important democratic institutions.[256] Concerns about the erosion of public discourse are not new; they have long followed concerns over the rise of social media generally.[257] These concerns were especially acute in the wake of the riots of January 6, 2021, which were propagated at least in part based on false information about the results of the 2020 presidential election.[258] The accounts surrounding these events have focused on fake news and misinformation. They involve the insidiousness of rumors and our biases towards conspiracy theories. But what share of the blame of our new digital ecosystem belongs to the fact that we live in a visual medium that is permeated with fakery?

Here is one plausible account. As people live more of their lives in spaces where they cannot know whether to trust the images they see, they will come to distrust the information that is shared. Under this narrative, the filter and the fake news story are part of the same information landscape, where everything is believable enough but also, possibly, faked. As the percentage of material that is faked goes up—and we know that most people are manipulating their online presentations of self—we

> G.P. would say, then what is science, and is it all-encompassing and altruistic and without error and always acting in the interests of humanity? These questions had been plaguing Goop for a while – not just what is a fact, or how important is a fact, but also what exactly is Goop allowed to be suggesting?

Id.

255. *Id.*

256. In discussing the more specific practice of "stealth marketing," Ellen Goodman argues that it "harms . . . by degrading public discourse and undermining the public's trust in mediated communication." Goodman, *supra* note 34, at 87.

257. *See* PARISER, *supra* note 33; *see also* Khiara M. Bridges, *Language on the Move: "Cancel Culture," "Critical Race Theory," and the Digital Public Sphere*, 131 YALE L.J. FORUM 767, 770–71 (2022). Bridges writes:

> On social media, rational debate – the hallmark of the civic deliberations that took place in the Habermasian public sphere – is not a dominant presence. When one dares to open the Twitter app, one is more likely to encounter abusive speech, ad hominem attacks, and wildly fact-free and logic-free statements than rational argumentation.

Id.

258. *See* Bill McCarthy, *Misinformation and the jan. 6 insurrection: when 'patriot warriors' were fed lies*, POLITIFACT (June 30, 2021), https://bit.ly/3MxY4w0.

should expect that people will distrust what they see, and that distrust will spill over into the other kinds of digital information they consume.

Consider again the "Moon Juice" peddler, Amanda Chantal Bacon. Her clients are made up of the likes of "Gwyneth Paltrow, Emma Roberts, and Shailene Woodley."[259] She is, famously, a critic of "Western medicine."[260] Chantal Bacon proudly notes that she has "never paid influencers."[261] Instead, "[s]ocial media, and specifically Instagram, has always been important to me."[262] That is, the very nature of the social media platforms allows her company to flourish. Instagram provides her with a space to promote her products for profit, by positioning herself as an expert on "'wellness and longevity'" and on leading a "'holistic lifestyle.'"[263]

Consider now Alex Jones who, like Chantal Bacon, is a salesman. He has a website, InfoWars, where he offers "organic fair-trade coffee" that "can be purchased in an 'Immune Support' variety that includes cordyceps and reishi mushroom extracts."[264] He also markets probiotics and a "Super Female Vitality" supplement. These are made from the exact same extracts that Moon Juice sells.[265] Jones is perhaps best known as a right-wing radio host and conspiracy theorist, who was instrumental in propagating the false narrative that the 2020 election was stolen from Donald Trump.[266] He spoke at a rally in DC on January 5, and was subpoenaed to discuss his knowledge of, and involvement in, the January 6 attack.[267]

The link between Chantal Bacon, Jones, and January 6, is not as far-fetched as it might initially seem. Chantal Bacon and Jones both traffic in, and profit from, misinformation, and their business models depend on a distrust of facts presented by sources external to themselves. The background conditions of the platforms make it so that truth becomes a commodity—with real world consequences that include harming our most time-worn democratic institutions.[268] This entwinement of profit and

259. Elana Lyn Gross, *How Moon Juice's Founder Built Her Wildly Popular Wellness Brand*, FORBES (Aug. 13, 2018), https://bit.ly/3QyIDr5.

260. *Id.*

261. Gross, *supra* note 259.

262. *Id.*

263. *Id.*

264. *See* Young, *supra* note 34.

265. *See id.*

266. *See* Frontline, *What Conspiracy Theorist Alex Jones Said in the Lead Up to the Capitol Riot*, PBS (Jan. 12, 2021), https://bit.ly/3oyZK0n.

267. *See* Benjamin Siegel, *Conspiracy theorist Alex Jones reveals he appeared before Jan. 6 committee*, ABC NEWS (Jan. 25, 2022, 6:54 PM), https://bit.ly/3oGYhVB.

268. *See* David Remnick, *The Devastating New History of the January 6th Insurrection*, THE NEW YORKER, (Dec. 22, 2022), https://bit.ly/47t5ASt (describing the attack as "a deliberate, coordinated assault on American democracy that could have easily ended with the kidnapping or assassination of senior elected officials, the emboldenment of extremist groups and militias, and, above all, a stolen election, a coup").

politics on social media disrupts the possibility of creating any shared civic discourse.[269]

The distrust of institutions that leads one to rely on the self—which must be improved, immunized, and reinforced with various supplements to withstand whatever may come—straddles the spectrum of political parties as it does socioeconomic class. As individuals' focus is pulled into themselves and their interests, it is pulled away from the creation of a collective reality or of a shared community.[270] The result is that people feel more alone and more siloed and therefore less engaged as members of society. As people spend more of their lives online, and occupy spaces that are rife with fakery, they will be less inclined to engage in a common civic life.[271]

V. PLATFORM REGULATION

We have argued for a more robust assessment of the costs of shallow fakes in both scholarship and policy. What does that mean in terms of regulation? The first, and most obvious, regulatory move is to demand greater transparency from social media platforms. Relatedly, the FTC should sharpen and expand its guidelines around deception on social media. Finally, we think there is room for voluntary initiatives by social media firms, akin to the work being done in countering violent extremism and child sexual abuse, though we note that some of the dominant policy proposals today—especially antitrust policies aimed at greater competition—are likely to be unhelpful in this context and may instead make the problem worse.

Our focus in this Part is on the platforms, not the users, because it is the platforms that are best situated to address the problem. They create the market for shallow fakes, they have the most information about what is happening on their services, and, crucially, they control users' experiences by incentivizing them to engage in shallow fakery. While platforms enjoy broad immunity for much of what their users do under Section 230 of the

269. As researcher Peter Dahlgren has noted in his work on social media and democracy, "from the standpoint of users, even if our intentions are civic or political, we are still addressed by and embedded in dominant online consumerist discourses." Dahlgren, *supra* note 242, at 2060–61 ("These discourses offer us subject positions mostly as consumers, rarely as citizens.").

270. *See* TOLENTINO, *supra* note 240, at 30 ("Facebook's goal of showing people only what they were interested in seeing resulted, within a decade, in the effective end of shared civic reality.").

271. Jia Tolentino's interview of author Naomi Klein discusses the consequences of maintaining this hierarchy of priorities. Klein explains: "[T]he amount of labor we are putting into optimizing our bodies, our image, our kids, is robbing from the work that needs to be done to preserve the habitability of the planet, to preserve our humanity in the face of those spasms." Jia Tolentino, *Naomi Klein Sees Uncanny Doubles in Our Politics*, THE NEW YORKER INTERVIEW (Sept. 10, 2023), https://bitly.ws/Uthu.

Communications Decency Act, they do so only if they are not "responsible, in whole or in part, for the creation or development of" the offending content.[272] We are centrally concerned here with what the platforms themselves do. Moreover, as the Ninth Circuit explained in *Fair Housing Council of San Fernando Valley v. Roommates.com, LLC*, "a website may be immune from liability for some of the content it displays to the public but be subject to liability for other content."[273]

A. Transparency Reforms

Before we are ready to prescribe platform regulations with any detail, we need to know much more about the platforms' internal operations. How are they offering filters and who are they targeting? How much are they filtering by default, without giving users adequate notice? Relatedly, how much user data are they tracking and towards what ends? We should also know more about user behavior and how it is shaped by the platforms' dark patterns. All of this to say: we simply do not know enough about what social media platforms are doing and how people are using them.

For many of the most important questions, only the social media platforms have the answers or have access to data that could provide answers. Unfortunately, the platforms have not been terribly transparent. As Facebook insider Frances Haugen testified, "I came forward because I recognized a frightening truth: almost no one outside of Facebook knows what happens inside Facebook."[274] Despite repeated calls to remedy this problem, Facebook has refused to share its internal research. As the Wall Street Journal reported last year, "Facebook has consistently played down the app's negative effects on teens, and hasn't made its research public or available to academics or lawmakers who have asked for it."[275] Tellingly, the first time most of the public and regulators became aware of the scope of the mental health crisis on social media was when a Facebook employee blew the whistle and leaked internal research.[276]

---

272. 47 U.S.C. §§ 230(c) and (f)(3) (1998).

273. Fair Hous. Council of San Fernando Valley v. Roommates.com, LLC, 521 F.3d 1157, 1162–63 (9th Cir. 2008) (finding that Section 230 did not immunize Roommates.com from liability under public accommodations laws for the actions the platform took, including actions that influenced user behavior). While the Supreme Court recently considered reviewing Section 230's scope, the Court ultimately sidestepped the issue. *See* Gonzalez v. Google LLC, 598 U.S. 617, 622 (2023) ("We therefore decline to address the application of § 230 to a complaint that appears to state little, if any, plausible claim for relief.").

274. *Testimony From A Facebook Whistleblower*, *supra* note 139.

275. *See* Wells et al., *supra* note 21.

276. *See* Scott Pelley, *Whistleblower: Facebook is misleading the public on progress against hate speech, violence, misinformation*, CBS NEWS: 60 MINUTES (Oct. 4, 2021), https://bit.ly/3og1Xh4.

While platforms publish transparency reports and have undertaken efforts to share their inner workings, these have been a limited and woefully inadequate substitute for independent research.[277] These transparency reports mostly focus on content takedowns and law enforcement requests; they largely ignore the problems created by the platforms' algorithms or filter tools.[278] When platforms do share information more broadly, they choose to do so only with hand-picked researchers.[279] When these efforts are more ambitious, like the Social Science One project, which aimed to give researchers insider access to social media data, they have stalled.[280]

On their own, then, platforms have done too little on transparency. Even worse, they have aggressively blocked and threatened independent researchers seeking to study what kinds of material people consume on social media. For example, when researchers at New York University put together a tool to study how political advertisements are targeted on the platforms—something that is of considerable public interest—Facebook sent them a cease-and-desist letter, ultimately shutting off their access to the platform for violating the terms of service.[281] Other researchers have had similar experiences.[282]

Given the current landscape, the following two reforms would make considerable inroads: (1) the platforms should be required to make more information available to independent researchers and the public at large; and (2) independent researchers should not fear criminal or civil liability for engaging in non-commercial, public-interest research. Outlining the full contours of a legislative solution to this problem is beyond the scope of this Article, but the Platform Transparency and Accountability Act is a good start.[283] That bill would create a safe harbor for researchers engaged

277. *See, e.g.*, Mathew Ingram, *Facebook "transparency report" turns out to be anything but*, COLUM. JOURNALISM REV. (Aug. 26, 2021), https://bit.ly/43dqz8Q.

278. For example, Meta's recently created "Widely Viewed Content Report" is supposed to bring transparency into the way information goes viral on the platform, yet it provides such a high-level overview, it gives only a snapshot of "what a typical feed looks like," and does not give granular information about how information spreads. *See, e.g.*, *Widely Viewed Content Report: What People See on Facebook* Q1 2023 Report, META, https://bit.ly/3OoHlxm (last visited Sept. 12, 2023).

279. *See* Steven Levy, *It's Time to Talk About Facebook Research*, WIRED (Sept. 17, 2021), https://bit.ly/42Whv8H.

280. *See* Craig Timberg, *Facebook made big mistake in data it provided to researchers, undermining academic work*, WASH. POST (Sept. 10, 2021), https://wapo.st/3WjbWi2.

281. *See* Taylor Hatmaker, *Facebook cuts off NYU researcher access, prompting rebuke from lawmakers*, TECHCRUNCH (Aug. 4, 2021, 8:17 PM), https://bit.ly/3DrXoEs.

282. *See* Ethan Zuckerman, *Facebook has a misinformation problem, and is blocking access to data about how much there is and who is affected*, THE CONVERSATION (Nov. 2, 2021, 8:27 AM), https://bit.ly/41MtJPW.

283. *See* Senator Coons Press Release, *Coons, Portman, Klobuchar Announce Legislation to Ensure Transparency at Social Media Platforms* (Dec. 9, 2021),

in noncommercial research and also allow the National Science Foundation to act as a kind of gatekeeper for reviewing research proposals and requiring platforms to provide relevant data.[284] Mandating this kind of platform transparency and protecting researchers seeking transparency are essential first steps to exposing the incidence and the impact of mass deception in our lives online.

B. Deceptive and Unfair Trade Practices by the Platforms

There are few legal restrictions that address shallow fakes. Where they exist, they are difficult to interpret and sparingly enforced. Overall, agency regulation of deception in social media has been thin, sporadic, and entirely sectoral. Kim Kardashian, one of the most successful Instagram influencers with hundreds of millions of followers, has been fined by a regulatory agency for deceptive marketing only once—by the Securities and Exchange Commission (SEC), in relation to cryptocurrency advertising, and never by the FTC.[285] This example illustrates a further limitation of existing regulation—that it targets individual users, rather than the social media platforms.

One obvious context where fakery clearly has the potential to be regulated is in advertisements. The rules that have been promulgated, however, tend to be tangential to the problem of shallow fakes, and hardly ever enforced. In 2019, the FTC promulgated an updated memo on the pre-existing guidelines for influencers on social media.[286] The guidelines include clarifications around when and how to disclose a *paid* partnership. They indicate, for instance, that "[i]f your endorsement is in a *picture* on a platform like Snapchat and Instagram Stories, superimpose the disclosure over the picture and make sure viewers have enough to notice and read it."[287] Additionally, the FTC asserts that "[y]ou can't talk about your experience with a product you haven't tried" and that influencers cannot say a product is great if they have tried it and thought it was

https://bit.ly/3pZgVZ3. For a more detailed account of why researchers need protected access to social media platform data, see Alex Abdo et al., *A Safe Harbor for Platform Research*, KNIGHT FIRST AMEND. INST. (Jan. 19, 2022), https://bit.ly/3MEPNaU.

284. *See* Jeff Horwitz, *Senators Want Social-Media Apps to Share Research*, WALL ST. J. (Dec. 9, 2021), https://on.wsj.com/3ojV1zt.

285. *See* Press Release, *SEC Charges Kim Kardashian for Unlawfully Touting Crypto Security*, U.S. SECURITIES AND EXCHANGE COMMISSION (Oct. 3, 2022), https://bit.ly/3Wrjj7d (describing $1.26 million fine for Kardashian's failure to disclose that she was paid $250,000 to post her support of crypto asset EthereumMax).

286. *See Disclosures 101 for Social Media Influencers*, FEDERAL TRADE COMMISSION (Nov. 2019), https://bit.ly/41STyxC; *see also* 16 CFR Part 255, *Guides Concerning the Use of Endorsements and Testimonials in Advertising*, FEDERAL TRADE COMMISSION (2009), https://bit.ly/3WirfHI.

287. *Disclosures 101 for Social Media Influencers*, *supra* note 286, at 4.

terrible.[288] These are helpful, to be sure. But these guidelines do not apply to the vast majority of the deception we describe, in which people are not actively promoting a specific product. Even in the narrow situations they are meant to cover, Alexandra Roberts has found that "[f]alse advertising claims based on the use of editing software to improve people's appearance" are unlikely to succeed because the FTC looks for explicitly misleading statements about a product's efficacy.[289] These rules are also poorly enforced.[290] As such, one study estimated that only 7% of all sponsored influencer posts comply with FTC rules.[291]

Because the FTC has not been especially active in this space, some scholars have turned to the Lanham Act as a promising option, given that it allows for causes of action to be privately enforced.[292] But making out a successful claim is still difficult. In *Lokai Holdings, LLC v. Twin Tiger USA, LLC*, Twin Tiger alleged, among other things, that competitor Lokai's "failure to disclose that it compensates certain influencers, celebrities, and media outlets for their endorsement of Lokai products in online and social media advertising is likely to deceive reasonable consumers."[293] Yet the district court denied the claim, concluding that "the Lanham Act does not impose an affirmative duty of disclosure."[294] Accordingly, "failure to disclose compensation to celebrities and influencers for promoting its products is not actionable under the Lanham Act."[295]

These examples address only one small slice of the fakery taking place online, and the focus is always on individual users or brands, as opposed to the platforms or their policies as a whole. Our reforms are aimed at the platforms themselves. Specifically, we call upon the FTC to promulgate rules in this context. While we are not the first to do so,[296] our

---

288. *Id.* at 6.

289. Roberts, *supra* note 21, at 114. Roberts writes:

> [T]hose that equate to a false or misleading statement about the product's efficacy—like photoshopping whiter teeth in an influencer ad for a tooth whitening product or longer lashes in an ad for a lash-lengthening mascara—seem more likely to be fair game.

Id.

290. *See id.* at 120.

291. *See* MEDIAKIX, *supra* note 15.

292. *See* Roberts, *supra* note 21, at 86–88 (noting that the FTC "lacks the resources and perhaps the authority to enforce industry-wide change" but that the Lanham Act allows for private companies to sue one another and arguing in favor of "private actors . . . us[ing] the Lanham Act to challenge competitors' false influencing").

293. Lokai Holdings, LLC v. Twin Tiger USA, LLC, 306 F. Supp. 3d 629, 639 (2018).

294. *Id.* at 640.

295. *Id.*

296. At least four student notes have identified this problem along with several practitioners. *See, e.g.*, Lauryn Harris, *Too Little, Too Late: FTC Guidelines on "Deceptive and Misleading" Endorsements by Social Media Influencers*, 62 HOW. L.J. 947 (2019);

emphasis is much broader than what has been previously suggested: we are concerned with holding the platforms, rather than individual users, influencers, or even brands, accountable.[297] As we have shown, the platforms are an unusual market, but a market nonetheless—they are advertising platforms where attention is being monetized.[298] Everything everyone posts on the platforms, even if not directly sponsoring a product or service, is in some sense advertising, contributing directly or indirectly to the promulgation of the advertising market run by the platforms. As such, it would not be an extreme step beyond previous exercises of FTC authority to take action to address the confusion, fraud, and mental health impact on platform consumers and users.

The FTC, under the authority of the Federal Trade Commission Act, has the necessary leeway to act.[299] We can imagine two specific actions: (1) new FTC rules around platforms that discourage deception among users, and (2) more enforcement by the FTC against platforms around these issues. The FTC was granted the authority, in part, to regulate deception in the market.[300] And the FTC has regulated other aspects of consumer welfare on these platforms, including information privacy.[301]

What should those rules contain? The FTC's own guidelines for influencers offer a starting point. The FTC's focus has been on disclosure, so that users know that someone peddling a product is in fact doing so. This is, unsurprisingly, precisely the opposite of what advertisers and the platforms want.[302] Thus, the FTC could similarly advocate for two

---

Laura E. Bladow, *Worth the Click: Why Greater FTC Enforcement Is Needed to Curtail Deceptive Practices in Influencer Marketing*, 59 WM. & MARY L. REV. 1123 (2018); Tisha James, *The Real Sponsors of Social Media: How Internet Influencers Are Escaping FTC Disclosure Laws*, 11 OHIO ST. BUS. L.J. 61 (2017); Christopher Terry et al., *Throw the Book at Them: Why the FTC Needs to Get Tough With Influencers*, 29 J. L. & POL'Y 406 (2021).

297. The problem has been framed as finding plaintiffs, *see* Roberts, *supra* note 21, at 87–88, but we are ultimately concerned with identifying the correct defendant. There are, of course, certain users, like influencers, and especially those with particularly large followings, who should be held accountable. The problems we have identified, however, are with *platforms'* policies, and our concern here is therefore with regulating *platform* behavior.

298. *See generally* WU, *supra* note 10.

299. *See* 15 U.S.C. § 45(a).

300. *See* F.T.C. v. Colgate-Palmolive Co., 380 U.S. 374, 384–85 (1965) (describing how the Federal Trade Commission Act was significantly amended in 1938 to include a prohibition on "deceptive acts or practices in commerce").

301. *See* Daniel J. Solove & Woodrow Hartzog, *The FTC and the New Common Law of Privacy*, 114 COLUM. L. REV. 583, 586 (2014) (describing how the FTC has built up an extensive body of regulations regarding privacy, which "is functionally equivalent to a body of common law").

302. *See* Roberts, *supra* note 21, at 84–85 (noting that "[o]mitting sponsorship disclosure enables paid content to masquerade as organic buzz and peer-to-peer testimonial, rendering misrepresentations even more persuasive," and so "the majority of influencers and brands go out of their way to obscure the nature of their relationship").

important kinds of disclosure. First, they could require platforms to clarify when content on the platform is being altered and how. If TikTok videos are filtered, the filtering should be plainly explained to users. These laws would be directed at the platforms to address some of the harms that stem from digital manipulation. Laws could include, for example: demanding transparency around manipulated videos and filters; requiring users to share whether their media has been manipulated; using automated tools to screen for manipulated media; discouraging the use of manipulated media; and monitoring users for mental health problems related to manipulated media (monitoring of the kind that Facebook was undertaking in secret).[303] Once the FTC knows more about the way the algorithms function, it could make specific recommendations based on the searches the users undertake and the algorithms that lead to specific results.

Second, users themselves could be encouraged to be transparent about the modifications they apply to their posts. While our focus is on the platforms, addressing the users might be an effective supplement and a way to enhance user autonomy. The users themselves could help change the baseline from the barely perceptible filter to the filter labeled as such, especially for those who have large followings.[304]

Once these rules are in place, the FTC should step up its enforcement. Enforcement under these new regulations has the potential to be less involved and less costly in at least one way: rather than seek out individual actors online, the FTC can focus its resources on the limited number of social media platforms that control all our online activity.

C. Other Initiatives

In the absence of new regulations, industry reforms are a second-best solution. While there are good reasons to be skeptical of voluntary industry initiatives, there are several compelling precedents of industry-wide norms developed by social media firms. For example, the Global Internet Forum for Countering Terrorism ("GIFCT") allows firms to harmonize their efforts to combat extremist imagery, along with other counter-terrorism steps. In 2017, the firms that make up the GIFCT created an industry database of "perceptual hashes of known images and videos produced by

303. *See facebook files*, *supra* note 25.

304. Users with large numbers of followers are also privy to an especially close relationship with the platforms. As Frier explains:

> Today, Kim Kardashian West has 157 million followers and makes about $1 million for a single post. Paris Hilton eventually joined Instagram too, and now has 11 million followers. Porch [who works at Instagram] now has employee counterparts in Los Angeles who answer celebrity queries for the Kardashians and others, solving their problems directly while most of the app's users fend for themselves.

FRIER, *supra* note 4, at 138–39.

terrorist entities on the United Nations designed terrorist groups lists—which GIFCT members had removed from their services."[305] Another example is the hash database that the firms share with the National Center for Missing and Exploited Children ("NCMEC") for child sexual abuse material ("CSAM").[306] This is a massive program, handling tens of millions of reports of CSAM in a year, and most of these are from large social media platforms.[307]

While the GIFCT and NCMEC hash databases are different in important respects—the former is an entirely private multistakeholder initiative while the latter is congressionally mandated—they offer lessons for the development of industry-wide initiatives. We could imagine the firms that make consumer products to advertise coming together in agreement—perhaps along with advertising agencies—to self-regulate in various ways. At a minimum, they could all endorse the FTC guidelines, which include rules around transparency and disclosure about the products being promoted. These firms could further agree not to use filters in their own posts, which could encourage everyday social media users to do the same. While this might sound far-fetched, it has happened in traditional media, where a considerable effort has been made to reduce the use of airbrushing and to encourage a wider acceptance of different body types.[308] For example, Olay, the cosmetics giant, has pledged to stop using airbrushing in its advertising campaigns.[309]

But we have already seen that the problem goes far beyond what are technically considered advertisements, so we would not want to limit these initiatives to firms seeking to advertise. It would be much more impactful to apply them to the platforms themselves. This would mean that Instagram (and its owner Meta), Snapchat, and TikTok, among others, would devise a set of limitations for modified content. They need not wholly ban filters and other popular methods of altering images—it would already be significant if they placed a label on images that were digitally altered, much in the way that advertisers have proposed noting which ads

305. David Cohen, *Global Internet Forum to Counter Terrorism Expands Scope of Its Database*, ADWEEK (July 26, 2021), https://bit.ly/44YHtsG (last visited Jul. 9, 2023).

306. *See* Anirudh Krishna, Note, *Internet.gov: Tech Companies as Government Agents and the Future of the Fight Against Child Sexual Abuse*, 109 CAL. L. REV. 1581, 1602 (2021).

307. *See id.* at 1603.

308. *See, e.g.*, Vanessa Friedman, *Airbrushing Meets the #MeToo Movement. Guess Who Wins*, N.Y. TIMES (Jan. 15, 2018), https://nyti.ms/3MI6ulM (describing plans by CVS and other firms to stop "materially altering" imagery in advertisements); *see also* Eric Pfanner, *A Move to Curb Digitally Altered Photos in Ads*, N.Y. TIMES (Sept. 27, 2009), https://bit.ly/3IrJzbG (describing a French law to ban airbrushing).

309. *See* Becky Bargh, *Olay pledges to stop airbrushing advert campaigns*, COSMETICS BUS. (Feb. 24, 2020), https://bit.ly/3BKC7ov.

have been significantly altered.[310] This labeling would allow users to better distinguish between real and fake images and give them the tools to more accurately interpret the posts.

This kind of initiative, which is ambitious on its own, would still not remedy many of the problems described here. Users could still voluntarily edit their photos using native photo-editing software and then upload them to social media platforms, which likely would not be able to identify whether the image was doctored. Even sophisticated face-recognition software depends on a training image; if someone only posts edited images of their face, that would be the face identified by the software. And many of the problems described here—like a photo simply being taken out of context—do not only concern editing software.

Any initiative, therefore, must also include media literacy training. For more robust protection against trickery, social media users need to be critical consumers who are versed in identifying fakes and, more importantly, taking what they see with a healthy (rather than democracy-defeating) dose of skepticism so that they do not think everything is fake and can instead assess the difference between fact and fiction. Educating users to be critical will not fundamentally change the nature of social media platforms,[311] but it will enable users to make a distinction between the digital and the physical world in ways that give users more control and more information about what exactly they are consuming.

One final way that industry norms could help would be to cultivate a diversity of images across users' visual fields. Just as people speak of needing a balanced "information diet" to combat the filter bubble, platforms could ensure that users receive a balanced diet of images—filtered and unfiltered.[312] This would go some way towards combatting the unrealistic standards that are presented as the norm on social media platforms. Being exposed to images of different body types would also put the brakes on the common experience of tunneling down a filter bubble where one is fed only one particular kind of image.

VI. CONCLUSION

There is a great deal of talk about the "metaverse," a digital world where we will be able to leave our bodies behind.[313] The foundations for

310. *See, e.g.*, Vanessa Friedman, *supra* note 308 (describing how CVS will mark images that are not significantly digitally altered).

311. *See* TOLENTINO, *supra* note 240, at 19 ("The internet is engineered for this sort of misrepresentation; it's designed to encourage us to create certain impressions rather than allowing those impressions to arise 'as an incidental by-product of [our] activity.'").

312. *See* Steven Leckart, *Balance Your Media Diet*, WIRED (Jul. 15, 2009), https://bit.ly/45bAvS6.

313. *See* Eric Ravenscraft, *What Is the Metaverse, Exactly?*, WIRED (Nov. 25, 2021), https://bit.ly/3MpVGHs.

that world are being built on today's digital platforms.[314] But today's platforms are awash in deceit, through deepfakes and shallow fakes alike. Ours is a visual field marked by intense pressure to conform to a particular ideal of the perfect self, one that can only ever exist in a fake world.[315] For most of us, though, our online images are not our reality. Let us keep it that way.

314. *See* Katherine Fung, *Facebook Changes Company Name to 'Meta' in Rebrand, Social Network Name Will Stay*, NEWSWEEK (Oct. 28, 2021, 2:38 PM), https://bit.ly/44DLyD5 ("The metaverse is the next evolution of social connection. Our company's vision is to help bring the metaverse to life, so we are changing our name to reflect our commitment to this future.").

315. Consider @lilmiquela who is "a 19-year-old Robot living in LA." *See @lilmiquela*, INSTAGRAM, https://bit.ly/3Iwl4dI (last visited June 26, 2023). Lucy Blakiston explains:

> Miquela's 'life' is the definition of unrealistic, because she's literally an amalgamation of a perfectly selected bunch of pixels. But between Photoshop, Facetune, filters and all the other bullshit that exists these days to help us airbrush our online lives, Miquela's façade really isn't that different from any other influencer. And maybe that's the point.

Lucy Blakiston, *Lil Miquela and the rise of the robot influencer*, THE SPINOFF (July 15, 2021), https://bit.ly/3r72ByA.