THE RULE OF LAW ON INSTAGRAM: AN EVALUATION OF THE MODERATION OF IMAGES DEPICTING WOMEN'S BODIES

ALICE WITT,* NICOLAS SUZOR** AND ANNA HUGGINS***

This article uses innovative digital research methods to evaluate the moderation of images that depict women's bodies on Instagram against the Western legal ideal of the rule of law. Specifically, this article focuses on the contested rule of law values of formal equality, certainty, reason giving, transparency, participation and accountability. Female forms are the focal point for our investigation due to widespread concerns that the platform is arbitrarily removing some images of women's bodies and, possibly, privileging certain body types. After examining whether 4944 like images depicting (a) Underweight, (b) Mid-Range and (c) Overweight women's bodies were moderated alike, we identify an overall trend of inconsistent moderation. Our results show that up to 22 per cent of images are potentially false positives – images that do not appear to violate Instagram's content policies and were removed. The platform's opaque moderation processes, however, make it impossible to identify whether images were removed by Instagram or by the user. This article concludes that the apparent lack of rule of law values in the platform's moderation processes, and

* Alice Witt is a PhD Candidate in the Digital Media Research Centre ('DMRC'), School of Law, Faculty of Law at the Queensland University of Technology ('QUT'), Brisbane, Australia (alice.witt@hdr.qut.edu.au). We sincerely thank the anonymous reviewers, Dr Daniel Joyce, participants of the Oxford Internet Institute's Summer Doctoral Programme 2018, Dr Kylie Pappalardo's HDR writing group at QUT, Dr Stefanie Duguay and Edward Hurcombe for their insightful feedback. We would also like to thank Edward Gosden and Lee Jones for their help with statistical analysis, and Nick Carey and the editors of the *UNSW Law Journal* for their outstanding editorial assistance. We are grateful to Professor Lyria Bennett Moses, Armin Alimardani and the other organisers of the UNSW Law, Technology and Innovation Junior Scholars Forum 2017, where an earlier version of this article was presented.

** Nicolas Suzor, PhD, is an Associate Professor in the Digital Media Research Centre, Faculty of Law at the Queensland University of Technology, Brisbane, Australia. A/Prof Suzor is the recipient of an Australian Research Council DECRA Fellowship (project number DE160101542).

*** Anna Huggins, PhD, is a Senior Lecturer in the School of Law, Faculty of Law at the Queensland University of Technology, Brisbane, Australia.

Instagram's largely unfettered power to moderate content, are significant normative concerns which pose an ongoing risk of arbitrariness for women and users more broadly. We propose ways that platforms can improve transparency, and advocate for the continued development of digital methods for empirical, legal analysis of platform governance. These improvements are crucial to help identify arbitrariness where it exists and to allay the suspicions and fears of users where it does not.

I INTRODUCTION

Users of online platforms are increasingly concerned about whether user-generated content is moderated in ways that are free from arbitrariness.[1] Content moderation refers to the processes through which platform executives and their moderators set, maintain and enforce the bounds of 'appropriate' content based on many factors, including platform-specific rules, cultural norms or legal obligations.[2] Decisions around the appropriateness of content are ultimately regulatory decisions in the way that they attempt to influence or control the types of content we see and how and when we see it.[3] The problem is that platforms moderate content within a 'black box' that obscures internal decision-making processes from the view of over two billion active monthly social media users around the globe.[4] The lack of transparency around the decisions that platforms make continues to limit public understandings of how user-generated content is moderated in practice.[5] In this

1 See, eg, Nicolas Suzor, 'Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms' (2018) 4(3) *Social Media + Society* 1; Jessica Anderson et al, 'Censorship in Context: Insights from Crowdsourced Data on Social Media Censorship' (Research Report, Onlinecensorship.org, 16 November 2016) <https://onlinecensorship.org/news-and-analysis/onlinecensorship-org-launches-second-report-censorship-in-context-pdf>; Ranking Digital Rights, '2018 Corporate Accountability Index' (Research Report, April 2018) <https://rankingdigitalrights.org/index2018/assets/static/download/RDRindex2018report.pdf>.

2 Alyssa Miranda, 'A Keyword Entry on "Commercial Content Moderators"' (2017) 2(2) *iJournal* 1; Sarah T Roberts, 'Content Moderation' in Laurie A Schintler and Connie L McNeely (eds), *Encyclopaedia of Big Data* (Springer, forthcoming 2019, copy on file with author) 1 <https://escholarship.org/uc/item/7371c1hf>.

3 Kate Klonick, 'The New Governors: The People, Rules, and Processes Governing Online Speech' (2018) 131 *Harvard Law Review* 1598, 1602; Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media* (Yale University Press, 2018).

4 Ranking Digital Rights, above n 1, 5–6; Marjorie Heins, 'The Brave New World of Social Media Censorship' (2014) 127 *Harvard Law Review* 325, 326; Sarah T Roberts, 'Digital Detritus: "Error" and the Logic of Opacity in Social Media Content Moderation' (2018) 23(3) *First Monday* <http://firstmonday.org/ojs/index.php/fm/article/view/8283/6649>; Facebook, *Stats* (31 December 2018) <https://newsroom.fb.com/company-info/>.

5 See, eg, Suzor, above n 1, 5.

context, there are increasing calls for empirical analysis that can help the public to better understand whether rules around content are enforced in ways that are free from arbitrariness, and to identify the real impacts that moderation can have on users as the subjects of platform governance.[6]

Ongoing controversies around the moderation of images that depict women's bodies on Instagram, a social media application ('app', or 'platform') for photo, video and message sharing,[7] underline how little is known about processes for moderating content in practice.[8] Some online news publications claim that Instagram, a subsidiary of Facebook, Inc. with over one billion monthly active users,[9] is 'removing' – also described as 'banning,' 'censoring' and 'deleting' – depictions of female forms in seemingly arbitrary or biased ways.[10] Among these claims is one that the platform is less likely to remove thin-idealised depictions of women.[11] These allegations of bias are concerning as they suggest that the platform is amplifying the expression of some female users while silencing others.[12] Such allegations are also

---

6 See, eg, Anderson et al, above n 1, 21–2; Ranking Digital Rights, above n 1; Solon Barocas and Andrew D Selbst, 'Big Data's Disparate Impact' (2016) 104 *California Law Review* 671, 731–2; Christian Sandvig et al, 'An Algorithm Audit' in Seeta Peña Gangadharan, Virginia Eubanks and Solon Barocas (eds), *Data and Discrimination: Collected Essays* (Open Technology Institute and New America, 2014) 6, 7–9.

7 In a recent publication, Instagram described itself as 'a social media app used to share photos, videos and messages': see Instagram and National PTA, *Know How to Talk with Your Teen about Instagram: A Parent's Guide*, 4 <https://www.pta.org/docs/default-source/files/events/backtoschool/parents-guide-to-instagram.pdf >.

8 See, eg, Sarah T Roberts, 'Aggregating the Unseen' in Arvida Byström and Molly Soda (eds), *Pics or It Didn't Happen* (Prestel, 2017) 17; Kasandra Brabaw, 'This Curvy Muslim Woman Is Speaking Out about Censorship on Instagram', *Refinery29* (online), 8 March 2017 <http://www.refinery29.com/2017/03/144177/curvy-muslim-censorship-instagram>; Nivi Shrivastava, 'Instagram Apologizes to Plus-Size Blogger for Removing Bikini Pics', *NDTV* (online), 10 June 2016 <http://www.ndtv.com/offbeat/instagram-apologizes-to-plus-size-blogger-for-removing-bikini-pics-1417571>.

9 Instagram, *Welcome to IGTV* (20 June 2018) <https://instagram-press.com/blog/2018/06/20/welcome-to-igtv/>.

10 Fox News, 'Fitness Blogger Hits Back after Instagram Removes Pic of Her Cellulite', *Fox News* (online), 27 February 2017 <http://www.foxnews.com/health/2017/02/27/fitness-blogger-hits-back-after-instagram-removes-pic-her-cellulite.html>; Shauna Anderson, 'Why Was THIS Photo Banned From Instagram? The Reason Will Make You Shake Your Head in Disbelief', *Mamamia* (online), 23 May 2014 <http://www.mamamia.com.au/photo-banned-from-instagram/>; Caroline Bologna, 'After Instagram Censored Her Photo, Mom Speaks Out about Body Image', *Huffington Post* (online), 24 November 2016 <http://www.huffingtonpost.com.au/entry/after-instagram-censored-her-photo-mom-speaks-out-about-body-image_us_57fff3cfe4b05eff5582968a>; Sarah Buchanan, 'Instagram Deleted This Photo of a Woman's Cellulite – But She Has Best Response', *Daily Star* (online), 4 March 2017 <http://www.dailystar.co.uk/fashion-beauty/593521/Cellulite-Instagram-deleted-photo-censorship>.

11 See, eg, Kashmira Gander, 'Body Hair and Sexuality: The Banned Photos Instagram Doesn't Want You to See', *The Independent* (online), 9 March 2017 <http://www.independent.co.uk/life-style/instagram-banned-photos-images-body-hair-sexuality-fat-race-feminism-archive-molly-soda-arvida-bystr-a7620101.html>.

12 See, eg, Onlinecensorship.org, *A Resource Kit for Journalists: Issue Areas* (September 2017) [1. The Human Body, Instagram] <https://onlinecensorship.org/content/a-resource-kit-for-journalists#Issue-Areas>.

surprising from a commercial point of view given that Instagram is particularly popular with women.[13] By contrast, some news publications show that thin-idealised images of women are also removed from Instagram, and claim that the platform is creating a positive space for the depiction of all body types.[14] However, a common complaint is that users do not know what rules apply to their content or why certain content is removed while other apparently similar content is not. These controversies raise important issues around the risk of arbitrariness in decisions around content, which is an ongoing cause for concern for all users of platform technology.

In response to calls for data that can shed light on content moderation in practice, this article empirically investigates whether images that depict women's bodies on Instagram are moderated in a way that aligns with Anglo-American rule of law values.[15] We argue in Part II that the rule of law is valuable as it institutionalises constraints on arbitrariness in the exercise of power across public and private domains.[16] We situate this article within the emerging, broader project of digital constitutionalism, which contends that public governance values can and should influence the private rules of non-state actors, including the policies of social media

13 Pew Research Center, *Appendix A: Detailed Table* (1 March 2018) <https://www.pewinternet.org/2018/03/01/social-media-use-2018-appendix-a-detailed-table/>; Stevie Chancellor et al, '#thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities' (Paper presented at the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing, San Francisco, 1 March 2016) 1202 <http://www.munmund.net/pubs/cscw16_thyghgapp.pdf>; Hannah Seligson, 'Why Are More Women than Men on Instagram?', *The Atlantic* (online), 7 June 2016 <https://www.theatlantic.com/technology/archive/2016/06/why-are-more-women-than-men-on-instagram/485993/>.

14 Ellie Cambridge, 'Model "Too Sexy for Instagram" Is Banned from the Social Media Site Days after Racy Sideboob Pics', *News.com.au* (online), 18 August 2017 <http://www.news.com.au/technology/online/social/model-too-sexy-for-instagram-is-banned-from-the-social-media-site-days-after-racy-sideboob-pics/news-story/f294df2147f5ba479416e2b2e3a00d18>; Maya Salam, 'Why "Radical Body Love" Is Thriving on Instagram', *The New York Times* (online), 9 June 2017 <https://www.nytimes.com/2017/06/09/style/body-positive-instagram.html>; Jennifer B Webb et al, 'Fat Is Fashionable and Fit: A Comparative Content Analysis of Fatspiration and Health at Every Size Instagram Images' (2017) 22 *Body Image* 53, 54.

15 It should be noted that the subjects of these images might not, in fact, identify as a 'woman' or 'female'. It is a limitation of the scope and method of this article that, by analysing decontextualized images against a binary classification of gender, we unfortunately are unable to sufficiently engage with the pressing concerns of transgendered and non-binary people.

16 Krygier posits that the *telos* of the rule of law is its opposition to arbitrary power, irrespective of the specific legal and institutional features that accompany it: see, eg, Martin Krygier, 'The Rule of Law: Legality, Teleology, Sociology' in Gianluigi Palombella and Neil Walker (eds), *Relocating the Rule of Law* (Hart Publishing, 2009) 45; Martin Krygier, 'Four Puzzles about the Rule of Law: Why, What, Where? And Who Cares?' in James E Fleming (ed), *Getting to the Rule of Law* (New York University Press, 2011) 64.

platforms.[17] Given that there is no universal set of rule of law values, this article focuses on formal equality, certainty, reason giving, transparency, participation and accountability,[18] which are well-established values in this Western democratic discourse.[19] We posit that any attempt by Instagram to moderate, or regulate, content should adhere to these basic rule of law safeguards.[20] Despite some well-founded critiques of formal rule of law values, which can, inter alia, replicate systemic social bias, we suggest that this ideal nonetheless provides useful language to name and work through some of the governance tensions between platforms and their users.

In Part III, we explain our black box method for empirically examining content moderation in practice when only parts of the system are visible from the outside.[21] As a discrete case study, we focus on whether like images that depict (a) *Underweight*, (b) *Mid-Range* and (c) *Overweight* women's bodies are moderated alike on Instagram. This is a topical case study given the previously mentioned concerns about potential arbitrariness in processes for moderating images depicting female forms and the relative dearth of empirical research into the platform to date.[22] Specifically, we develop and apply a black box method to examine 4944 like images of (a) *Underweight*, (b) *Mid-Range* and (c) *Overweight* women's bodies, none of which appear to be explicitly prohibited by Instagram's policies. After programmatically collecting images, we use content analysis to identify whether coded images in like categories of content were removed. We use this coding to investigate true negatives (images that do not appear to violate Instagram's policies and were not removed), and potential false positives (images that do not appear to

17 Suzor, above n 1, 1; Lex Gill, Dennis Redeker and Urs Gasser, 'Towards Digital Constitutionalism? Mapping Attempts to Craft an Internet Bill of Rights' (Research Publication No 2015-15, Berkman Center for Internet & Society, 9 November 2015) 2.

18 See, eg, Tom Bingham, *The Rule of Law* (Penguin Books Limited, 2011); Richard H Fallon Jr, '"The Rule of Law" as a Concept in Constitutional Discourse' (1997) 97 *Columbia Law Review* 1, 8. See further the discussion in Part II.

19 Jeremy Matam Farrall, *United Nations Sanctions and the Rule of Law* (Cambridge University Press, 2007) 40–1.

20 ACLU Foundation of Northern California et al, *The Santa Clara Principles on Transparency and Accountability in Content Moderation* (7 May 2018) <https://newamericadotorg.s3.amazonaws.com/documents/Santa_Clara_Principles.pdf>; Electronic Frontier Foundation et al, *Manila Principles on Intermediary Liability* (24 March 2015) <https://www.manilaprinciples.org/>.

21 See, eg, Maayan Perel and Niva Elkin-Koren, 'Black Box Tinkering: Beyond Disclosure in Algorithmic Enforcement' (2017) 69 *Florida Law Review* 181; Nicholas Diakopoulos, 'Algorithmic Accountability' (2015) 3 *Digital Journalism* 398, 404.

22 See, eg, Tim Highfield and Tama Leaver, 'Instagrammatics and Digital Methods: Studying Visual Social Media, from Selfies and GIFs to Memes and Emoji' (2016) 2 *Communication Research and Practice* 47.

violate Instagram's policies and were removed), in each of these categories.[23] We explain how the limited information that the platform provides about its processes for moderating content imposes constraints on our ability to draw definitive conclusions about whether an image was removed by the platform or by a user. By developing and testing a black box study, we are able to provide new insights about the limitations of Instagram's moderation processes for this type of analysis. These limitations stand in stark contrast to what we identify as industry best practice and the demands of stakeholders, including researchers and civil society groups, for greater transparency and accountability from platforms.

Next, in Part IV, we discuss the results and findings from this study. In contrast to Anglo-American rule of law ideals, our results show that images have been inconsistently moderated across all thematic categories. The odds of removal for an image that depicts an *Underweight* and *Mid-Range* woman's body is 2.48 and 1.59 times higher, respectively, than for an image that depicts an *Overweight* woman's body. Across these categories, we find that up to 22 per cent of images that were removed by Instagram or by the user do not breach the platform's policies, and are therefore potentially false positives. We explore some of the possible explanations for these inconsistent outcomes, noting the limits of independent verification given that Instagram conceals its regulatory system from public scrutiny. Overall, our results raise concerns around the alignment between Instagram's governance practices and Western rule of law values. We argue that the lack of formal equality, certainty, reason giving and user participation, and Instagram's largely unfettered power to moderate content with limited transparency and accountability, are significant normative concerns which pose an ongoing risk of arbitrariness for women and users more broadly.[24]

As we discuss in Part V, our results suggest that concerns around the risk of arbitrariness in the outcomes of content moderation might not be unfounded. The inconsistency in the probability of removal for different categories highlights an apparent problem that warrants further in-depth, substantive investigation, and we outline different options for ongoing research. This research continues the important

---

23 We determined whether images are potential true negatives or false positives based on Instagram's Terms of Use and Community Guidelines: Instagram Help Centre, *Terms of Use* (19 April 2018) <https://help.instagram.com/581066165581870?helpref=page_content>; Instagram Help Centre, *Community Guidelines* <https://help.instagram.com/477434105621119?helpref=page_content>. As explained further below, we are only able to estimate a maximal false positive rate that includes both removals by Instagram and by the individual posters themselves.

24 It should be noted that marginalised individuals and groups generally face a higher risk of arbitrariness in moderation when participating in online platforms: see, eg, Stefanie Duguay, Jean Burgess and Nicolas Suzor, 'Queer Women's Experiences of Patchwork Platform Governance on Tinder, Instagram, and Vine' (2018) *Convergence: The International Journal of Research into New Media Technologies* 1 <https://journals.sagepub.com/doi/10.1177/1354856518781530>. While we acknowledge this ongoing risk, empirically examining the moderation of content posted by minority users is outside the scope of this article.

work of developing the application of digital methods for empirical, legal analysis of the internal workings of Instagram and other platforms. Empirical evidence is key to not only better understanding content moderation in practice, but also users' and, indeed, society's ability to hold platforms to account for their governance decisions. We conclude this article with a call for greater transparency and accountability around platform governance, including greater certainty around the rules that apply to content and the provision of public reasons for content removal. We suggest that enhanced protections in this regard are crucial to help independently monitor the performance of moderation systems and address public concerns about potential arbitrariness, including systemic bias.

II THE VALUES OF THE RULE OF LAW

The rule of law provides an ideal of governance that is entrenched in the liberal, Anglo-American constitutional tradition,[25] and characterised by its opposition to arbitrary power.[26] We argue that the rule of law provides a useful conceptual lens for evaluating the ways that content is moderated or regulated on Instagram, as part of platform governance more broadly. Here it is useful to clarify the distinction between the concepts of 'governance' and 'regulation'. As Burris, Kempa and Shearing note, governance refers to the 'management of the course of events in the social system'.[27] Regulation can be conceived as that large subset of governance that is concerned with standard-setting, monitoring and enforcement.[28] We argue that moderation is a form of regulation over users, and that the problems of content moderation are problems of governance. We explain how Instagram appears to regulate content in the following Part.

Anglo-American scholars traditionally conceptualise governance and, indeed, the rule of law, in terms of the exercise of power by the state over its citizens.[29] This raises the important question of whether such a discourse should apply in the context of social media platforms, which are privately owned and governed. According to a strict legal categorisation, the law that governs the relationship between the state and

---

25 See, eg, Krygier, 'The Rule of Law: Legality, Teleology, Sociology', above n 16, 45 ff; Fallon, above n 18, 1; see Bingham, above n 18, 3, 9.

26 Martin Krygier, 'Transformations of the Rule of Law: Legal, Liberal, and Neo-' (Paper presented at KJuris Workshop, Dickson Poon School of Law, King's College London, 1 October 2014).

27 Scott Burris, Michael Kempa and Clifford Shearing, 'Changes in Governance: A Cross-Disciplinary Review of Current Scholarship' (2008) 41 *Akron Law Review* 1, 9, citing Scott Burris, Peter Drahos and Clifford Shearing, 'Nodal Governance' (2005) 30 *Australian Journal of Legal Philosophy* 30, 30.

28 Colin Scott, 'Analysing Regulatory Space: Fragmented Resources and Institutional Design' [2001] (Summer) *Public Law* 329, 341–5.

29 See, eg, Joseph Raz, *The Authority of Law: Essays on Law and Morality* (Oxford University Press, 1979) 212.

its citizens is 'public', while law that governs the relationship between parties is 'private'.[30] In exchange for access to a platform, users must agree to abide by the terms of service, which is a private consumer contract between platform owners and users.[31] From this perspective, especially in the United States ('US') where most online platforms are based, the constitutional discourse of the rule of law has almost no application in the private sphere of contractual agreements between parties.[32]

However, such rigid distinctions between public and private governance are increasingly difficult to sustain and justify in 'decentralised',[33] 'networked'[34] or 'pluralised'[35] regulatory environments. In this article, we follow Krygier's explicitly teleological approach to the rule of law, which underlines that this discourse has purchase across realms, contexts and actors.[36] The rule of law is valuable as it institutionalises constraints on arbitrariness in the exercise of power, irrespective of the specific legal and institutional features that accompany it.[37] Krygier calls into question the conventional assumption that threats of arbitrariness with which the rule of law is concerned are a state monopoly in the context of a society 'full of networks, nodes, fields, and orderings that have power over people in and around them'.[38] In other words, if we are concerned with addressing the risk of arbitrariness in the exercise of power, it should not matter whether the source of that power is public or private.[39] This approach is persuasive as it recognises that states are not the only actors that exercise power with public consequences and in ways that can potentially harm individuals or groups.

A growing body of literature recognises that non-state actors, like Instagram, have become 'the new governors'[40] of the digital age.[41] Contractual terms of service

---

30 See, eg, Michel Rosenfeld, 'Rethinking the Boundaries between Public Law and Private Law for the Twenty First Century: An Introduction' (2013) 11 *International Journal of Constitutional Law* 125, 125.

31 Nicolas Suzor, 'The Role of the Rule of Law in Virtual Communities' (2010) 25 *Berkeley Technology Law Journal* 1817, 1820.

32 Suzor, 'Digital Constitutionalism', above n 1, 3.

33 See, eg, Julia Black, 'Decentring Regulation: Understanding the Role of Regulation and Self-Regulation in a "Post-regulatory" World' (2001) 54 *Current Legal Problems* 103.

34 See, eg, Clifford Shearing and Jennifer Wood, 'Nodal Governance, Democracy, and the New "Denizens"' (2003) 30 *Journal of Law and Society* 400, 408.

35 See, eg, Christine Parker, 'The Pluralization of Regulation' (2008) 9 *Theoretical Inquiries in Law* 349.

36 Martin Krygier, 'Why the Rule of Law Is Too Important to Be Left to Lawyers' [2013] (4) *Law of Ukraine: Legal Journal* 18, 20–1.

37 See, eg, Krygier, 'The Rule of Law: Legality, Teleology, Sociology', above n 16, 45; Krygier, 'Four Puzzles about the Rule of Law: Why, What, Where? And Who Cares?' above n 16, 66.

38 Martin Krygier, 'The Rule of Law: Pasts, Presents, and Two Possible Futures' (2016) 12 *Annual Review of Law and Social Science* 199, 221.

39 Ibid.

40 Klonick, above n 3, 1598.

41 See, eg, Lawrence Lessig, *Code and Other Laws of Cyberspace* (Basic Books, 1999) 220; James Grimmelmann, 'Regulation by Software' (2005) 114 *Yale Law Journal* 1719; Colin Scott, 'Regulation in the Age of Governance: The Rise of the Post-regulatory State' in Jacint Jordana and David Levi-Faur (eds), *The*

arguably function as types of constitutional documents in the way that they establish the power of platform owners to regulate user-generated content, and set standards of 'appropriate' behaviour.[42] As Facebook CEO Mark Zuckerberg acknowledged in 2009, '[o]ur terms aren't just a document that protects our rights; it's the governing document for how the service is used by everyone across the world'.[43] Yet terms of service make poor constitutional documents. Unlike traditional constitutions, this contractual bargain affords platform owners 'complete discretion to control how the network works and how it used'[44] by users, who are the subjects of platform governance.[45] Instagram's Terms of Use, for instance, states that '[w]e can remove any content or information you share on the Service if we believe that it violates these Terms of Use, our policies (including our Instagram Community Guidelines), or we are permitted or required to do so by law'.[46] Users have little say in determining the content of terms of service, and there is little effective choice in the market – over two billion of the world's active, monthly social media users can either 'take it or leave it'.[47]

There is widespread unease over the risks, including potential arbitrariness, posed by platform governance.[48] The last twenty years of massive growth in internet services has given rise to serious concerns that the reality of internet governance continues to replicate and entrench systemic issues, including social bias and discrimination.[49] In response, scholars and non-governmental organisations, among

---

Politics of Regulation: Institutions and Regulatory Reforms for the Age of Governance (Edward Elgar Publishing, 2004) 145.

42 Suzor, 'Digital Constitutionalism', above n 1, 2.

43 Adweek Staff, 'Facebook Reverts Terms of Service after Complaints', *Adweek* (online), 18 February 2009 <http://www.adweek.com/digital/facebook-reverts-terms-of-service-after-complaints/>. Mark Zuckerberg has said, '[i]n a lot of ways, Facebook is more like a government than a traditional company': David Kirkpatrick, 'The Facebook Defect', *TIME* (online), 12 April 2018 <http://time.com/5237458/the-facebook-defect/>. See also Kyle Langvardt, 'Regulating Online Content Moderation' (2018) 106 *Georgetown Law Journal* 1353, 1357.

44 Suzor, 'Digital Constitutionalism', above n 1, 3.

45 Ibid. See also Kate Crawford and Tarleton Gillespie, 'What Is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint' (2016) 18 *New Media & Society* 410, 412; Rebecca Tushnet, 'Power without Responsibility: Intermediaries and the First Amendment' (2008) 76 *George Washington Law Review* 986, 987.

46 Instagram Help Centre, *Terms of Use*, above n 23, [Content Removal and Disabling or Terminating Your Account].

47 Suzor, above n 1, 6; Facebook, above n 4.

48 See, eg, David Kaye, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, 38th sess, Agenda Item 3, UN Doc A/HRC/38/35 (6 April 2018); David Kaye, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, 32nd sess, Agenda Item 3, UN Doc A/HRC/32/38 (11 May 2016) 4.

49 See, eg, Molly Dragiewicz et al, 'Technology Facilitated Coercive Control: Domestic Violence and the Competing Roles of Digital Media Platforms' (2018) 18 *Feminist Media Studies* 609; Nicolas Suzor et al,

others, are calling for mechanisms to protect individuals in the online environment.[50] Many of these calls stem from a belief that the '[rule of law] is a universal human good'.[51] For example, Tim Berners-Lee, the inventor of the World Wide Web, advocates for a 'Magna Carta of the Internet', which considers how we should constitute online social spaces and articulates a set of limits on private power in cyberspace.[52] Berners-Lee builds on earlier initiatives such as the Internet Rights & Principles Coalition's Charter and Rebecca MacKinnon's call to 'Take Back the Internet!'[53] A number of detailed analyses, which highlight the effects that governance by platforms can have on a wide range of issues including censorship and privacy, support these rallying calls.[54]

These calls raise the question: in what ways can rule of law values enhance platform governance? Opposing arbitrary power, which is the primary purpose, or *telos*, of the rule of law,[55] has been regarded as integral to warding off tyranny throughout centuries of political and legal thought,[56] and is very much alive in contemporary Western debates about the dangers of lawless, capricious or unchecked governing power.[57] Arbitrariness occurs when power is exercised unpredictably, or

---

'Human Rights by Design: The Responsibilities of Social Media Platforms to Address Gender-Based Violence Online' (2019) 11 *Policy & Internet* 84.

50 See, eg, Dynamic Coalition on Platform Responsibility, *Recommendations on Terms of Service & Human Rights* (November 2015) <https://www.intgovforum.org/cms/documents/igf-meeting/igf-2016/830-dcpr-2015-output-document-1/file>.

51 Brian Z Tamanaha, *On the Rule of Law: History, Politics, Theory* (Cambridge University Press, 2004) 137; E P Thompson, *Whigs and Hunters: The Origin of the Black Act* (Allen Lane, 1975) 266.

52 'Tim Berners-Lee Calls for Internet Bill of Rights to Ensure Greater Privacy', *The Guardian* (online), 28 September 2014 <https://www.theguardian.com/technology/2014/sep/28/tim-berners-lee-internet-bill-of-rights-greater-privacy>; Berners-Lee recently suggested a 'contract for the web': Alex Hern, 'Tim Berners-Lee on 30 Years of the World Wide Web: "We Can Get the Web We Want"', *The Guardian* (online), 13 March 2019 <https://www.theguardian.com/technology/2019/mar/12/tim-berners-lee-on-30-years-of-the-web-if-we-dream-a-little-we-can-get-the-web-we-want>.

53 See, eg, Internet Rights & Principles Coalition, 'The Charter of Human Rights and Principles for the Internet' (Booklet, 5th ed, January 2018) 8 <http://internetrightsandprinciples.org/site/wp-content/uploads/2018/10/IRPC_english_5thedition.pdf>; Rebecca MacKinnon, 'Let's Take Back the Internet!' (Speech delivered at TEDGlobal, Edinburgh, July 2011) <https://www.ted.com/talks/rebecca_mackinnon_let_s_take_back_the_internet/transcript?language=en#t-6750>.

54 See, eg, Ranking Digital Rights, above n 1; Anderson et al, above n 1, 4–7; Gennie Gebhart, *Who Has Your Back? Censorship Edition 2018* (31 May 2018) Electronic Frontier Foundation <https://www.eff.org/who-has-your-back-2018>.

55 See, eg, Krygier, 'The Rule of Law: Legality, Teleology, Sociology', above n 16.

56 See, eg, Krygier, 'Why the Rule of Law Is Too Important to Be Left to Lawyers', above n 36, 20; Suzor, 'Digital Constitutionalism', above n 1, 6.

57 See, eg, Krygier, 'The Rule of Law: Pasts, Presents and Two Possible Futures', above n 38, 199–200; David Mednicoff, 'Trump May Believe in the Rule of Law, Just Not the One Understood by Most American Lawyers', *The Conversation* (online), 5 June 2018 <http://theconversation.com/trump-may-believe-in-the-rule-of-law-just-not-the-one-understood-by-most-american-lawyers-97757 >.

when it is exercised in a way that takes no account of the perspectives and interests of affected parties.[58] In Parts IV and V, we use a black box method to determine whether processes for moderating like images that depict women's bodies on Instagram align with our selected rule of law values. We find that there is a lack of consistency and predictability in these regulatory processes, underscoring the need for mechanisms for reducing the risk of arbitrary decision-making.

As previously noted, rule of law values have the potential to serve as institutionalised constraints on arbitrariness in the exercise of power. It is important to note that there is, of course, no universal set of rule of law values,[59] which can differ markedly depending on traditions and conceptions of this ideal of governance.[60] In this article, we focus on the values of formal equality, certainty, reason giving, transparency, participation and accountability, which are well-established values in the Anglo-American constitutional tradition,[61] and reflect recurring themes and concerns about the moderation of images that depict women's bodies in practice. We argue that any attempt to moderate, or regulate, content on Instagram should reflect these rule of law values, which are undergirded by two central tenets of the Western ideal of the rule of law. The first is that all societal actors, including state and non-state actors and individuals, should be ruled by and obey the law.[62] The second is that the law should be capable of effectively guiding individual action so that all societal actors have an appreciation of their position in a legal system.[63] In the remainder of this section, we examine each of the rule of law values of formal equality, certainty, reason giving, transparency, participation and accountability in more depth. This informs the following analysis, in which we evaluate the extent to which our selected rule of law values appear to be evident in Instagram's current moderation processes, and make suggestions to address any gaps in the available procedural safeguards.

The first of the rule of law values that is a focus of this article is equality. This value can take the form of a formal obligation,[64] which principally requires the consistent treatment of individuals in like circumstances – that is, treating like cases

---

58 Krygier, 'Why the Rule of Law Is Too Important to Be Left to Lawyers', above n 36, 34.

59 See, eg, Jeremy Waldron, 'Is the Rule of Law an Essentially Contested Concept (in Florida)?' (2002) 21 *Law and Philosophy* 137.

60 See, eg, Farrall, above n 19, 40–1.

61 Ibid; see generally Bingham, above n 18; Suzor, 'Digital Constitutionalism' above n 1. See further the discussion in Part II.

62 Raz, above n 29, 213–4.

63 Ibid.

64 See, eg, Jonathon W Penney, 'Virtual Inequality: Challenges for the Net's Lost Founding Value' (2012) 10 *Northwestern Journal of Technology and Intellectual Property* 209.

alike.[65] Formal equality prohibits different treatment, or discrimination,[66] of any kind between persons based on 'ethnicity, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status'.[67] Equality can also take the form of a substantive obligation,[68] which is concerned with the results of the application of the law to individuals, rather than the consistent treatment of individuals alone.[69] Substantive equality might require inconsistent treatment of persons to achieve certain outcomes, including emphasising minority voices.[70] In the context of Instagram, we employ a formal conception of equality, based on the premise that content moderation should not be selective.[71] We argue that the platform should seek to ensure, to the extent possible, that images in like categories of content are moderated alike and in a way that is predictable rather than arbitrary.

Scholars have articulated a number of critiques of formal equality, especially from a feminist perspective.[72] One is that the Anglo-American rule of law tradition holds the white male archetype as the universal standard for consistent treatment.[73] Another is that simple measures of formal equality are not sufficient to articulate more substantive and complex concerns that undergird the experiences of users whose content is removed.[74] Whilst we acknowledge the pertinence of critiques of formal equality, we argue that this value remains an appropriate normative aspiration for our initial exploratory study. We proceed first in formal terms to start to test whether there is support for some users' claims, which we detail in the following Part, and to shed light on the platform's systems for moderating content more broadly. In doing so, we aim to lay the foundations for more substantive work. Thus, examining substantive equality for women in Instagram's processes of content

---

65 See Bingham, above n 18, 55–6; Catharine Barnard and Bob Hepple, 'Substantive Equality' (2000) 59 *Cambridge Law Journal* 562.

66 Penney, above n 64, 217; see Bingham, above n 18, 56; *International Covenant on Civil and Political Rights,* opened for signature 16 December 1966, 999 UNTS 171 (entered into force 23 March 1976) arts 2.1, 26.

67 Internet Rights & Principles Dynamic Coalition, *The Charter of Human Rights and Principles for the Internet* (August 2014) United Nations Human Rights Office of the High Commissioner <https://www.ohchr.org/Documents/Issues/Opinion/Communications/InternetPrinciplesAndRightsCoalition.pdf> 14.

68 Scholars have distinguished between cases where it is normatively important to consistently enforce the rules of an online social environment in an equal way from online environments that are designed to be arbitrary or biased (as in some games) or where the environment is designed to be limited to addressing the needs of particular communities: see Nicolas Suzor, 'Order Supported by Law: The Enforcement of Rules in Online Communities' (2012) 63 *Mercer Law Review* 523, 537; Bingham, above n 18, 55.

69 Barnard and Hepple, above n 65, 564.

70 Robin L West, *Re-imagining Justice: Progressive Interpretations of Formal Equality* (Ashgate Publishing Limited, 2003) 107.

71 Farrall, above n 19, 41.

72 See, eg, Patricia A Cain, 'Feminism and the Limits of Equality' (1990) 24 *Georgia Law Review* 803, 804.

73 See, eg, Denise Schaeffer, 'Feminism and Liberalism Reconsidered: The Case of Catharine MacKinnon' (2001) 95 *American Political Science Review* 699, 700–1.

74 See, eg, Iris Marion Young, *Justice and the Politics of Difference* (Princeton University Press, 1990).

moderation is outside the scope of this article, but remains an important topic for future research.

The rule of law values of certainty and reason giving can help to address some of the concerns around formal equality. At a minimum, we argue that certainty requires that rules around content are open and clear. Rules are arguably open when users are able to identify all of the rules, terms or guidelines that apply to content, and clear when there is a small 'penumbra of uncertainty' around the meaning of rules, terms or guidelines.[75] Certainty on Instagram could be promoted by using less ambiguous language in its terms and guidelines, and providing examples of the types of content that are prohibited and not prohibited, as well as copies of the internal guidelines that moderators follow.[76] The practice of giving reasons for decisions similarly promotes equal and predictable treatment of analogous subject matter. As Esty observes, '[t]he rationality of a policy choice can best be evaluated when it is written down, explained, and published'.[77] In terms of Instagram, we suggest that reason giving requires that each user whose content has been removed is notified about the reasons upon which this decision that affects their expression was made. Certainty and reason giving have the potential to contribute to the rule of law on Instagram by ensuring that content moderation is stable enough to guide the decisions and behaviours of users.

Transparency and participation are also a focus of this article. The concept of a black box is by definition opaque and secretive, whereas transparency describes a state of openness. As Colomer notes, '[t]ransparency is concerned with the quality of being clear, obvious and understandable without doubt or ambiguity'.[78] In the context of content moderation on Instagram, we argue that transparency requires that the platform's processes for moderating content and decision-making be as open as possible, with the reasons for moderating content clearly expressed in notice to users. Industry best practice suggests that platforms should publish a regular report that details, inter alia, how much content is removed, who removes content, for what and by what means.[79] Avenues for public participation enable users to engage with the information that a platform has made transparent, and articulate their views if a decision affects their interests.[80] Participation can encompass a number of activities and procedures through which the public can express their views, engage with the

---

75 H L A Hart, 'Positivism and the Separation of Law and Morals' (1958) 71 *Harvard Law Review* 593, 607; Raz, above n 29, 214.

76 See, eg, ACLU Foundation of Northern California et al, above n 20.

77 Daniel C Esty, 'Good Governance at the Supranational Scale: Globalizing Administrative Law' (2006) 115 *Yale Law Journal* 1490, 1529.

78 *Belgium v Commission* (C-110/03) [2005] ECR I-2801, [44] (Advocate General Ruiz-Jarabo Colomer).

79 See, eg, ACLU Foundation of Northern California et al, above n 20.

80 Esty, above n 77, 1530.

decision-makers that set, maintain and enforce rules around content, and increase their general awareness of the systems that moderate their content. However, there are limitations related to the proposed rule of law values, particularly the transparency ideal.[81] The value of transparency reporting by platforms is of course contingent on the quality of the transparent information and, even if a black box is opened, transparency and participation may not be sufficient to generate change. These rule of law values are therefore arguably better understood as components of broader mechanisms for accountability.

The rule of law value of accountability underlines the importance of moving beyond the mere provision of information. As Black notes, accountability can be defined in terms of the ways in which one actor 'gives account and another [actor] has the power or authority to impose consequences as a result'.[82] Accountability requires clarity around the identity of the decision-maker and steps in the decision-making process, which stands in stark contrast to black box processes in which the identity of the decision maker is not revealed, and the process by which decisions are reached occurs behind closed doors.[83] In relation to Instagram, we are concerned with vertical accountability between platforms and their users. Accountability in this context requires users to be able to make demands to prevent, redress or challenge the results of particular action or inaction. The contractual relationship between users and platforms is a significant barrier to users being able to hold platforms to account for the ways that they set, maintain and enforce rules that govern user-generated content. We are also concerned with a second dimension of accountability that relates to the role of regulators, non-governmental organisations and other external stakeholders in holding Instagram to account for its governance practices. One example of potential external accountability is provided by international human rights law; however, the primary responsibility to respect, protect and fulfil international human rights falls on states,[84] not platform owners. Accordingly, the prospects of Instagram being held to account under this framework if its decisions negatively affect users' self-expression and autonomy are limited. New methods, like the one employed in this study, are particularly useful here as they provide alternative means of holding decision-makers to account.[85] Enhanced access to data through these methods can lead to, inter alia, collaborations between stakeholders in

---

81 See especially Mike Ananny and Kate Crawford, 'Seeing without Knowing: Limitations of the Transparency Ideal and its Application to Algorithmic Accountability' (2018) 20 *New Media & Society* 973.

82 Julia Black, 'Constructing and Contesting Legitimacy and Accountability in Polycentric Regulatory Regimes' (2008) 2 *Regulation & Governance* 137, 150.

83 Esty, above n 77, 1528–9.

84 See, eg, Daniel Joyce, 'Internet Freedom and Human Rights' (2015) 26 *European Journal of International Law* 493.

85 See generally Nicolas Suzor et al, 'What Do We Mean When We Talk about Transparency? Towards Meaningful Transparency in Commercial Content Moderation' (2019) forthcoming *International Journal of Communication* <https://eprints.qut.edu.au/126386/2/126386.pdf>.

platform governance, including academics, journalists and civil society, who can potentially exert pressure on platforms to improve their governance practices.

In sum, we argue that rule of law ends will be promoted if Instagram's governance processes adhere to the basic constitutional safeguards of formal equality, certainty, reason giving, transparency, participation and accountability. The rule of law framework in this article provides a well-established language to start to name and work through what is at stake for women and other users in Instagram's potentially arbitrary exercise of power. More broadly, our rule of law framework, and the innovative black box method that we employ, are potentially transferable to other controversies on social media platforms where concerns about arbitrary moderation processes arise. As we show in the following Part, there are ongoing concerns about the extent to which processes for moderating content on Instagram reflect our selected rule of law values.

III INSTAGRAM MODERATES CONTENT BEHIND CLOSED DOORS

In this section, we outline how Instagram makes important regulatory decisions about user-generated content. To contextualise this discussion, we first provide some background on Instagram and its moderation processes. As previously noted, Instagram is a social media platform,[86] which is available as an app for Apple iOS and Android operating systems, and for the web. Instagram's Terms of Use states that the platform's mission, which pivots around the rhetoric of openness and connectivity, is '[t]o bring you closer to the people and things you love'.[87] Once a user uploads an image, they can add a filter and caption field – a brief explanation – with hashtags or emojis.[88] A hashtag is a metadata label that features a hash character (#) before one or more words.[89] Users can include up to 30 hashtags, like #fitgirl, in a caption, and a search for a hashtag will display images or video ('posts') that users have tagged with that hashtag.[90] Other users can react to a post through comments,

86 Instagram and National PTA, above n 7, 7.

87 Instagram Help Centre, *Terms of Use*, above n 23 [The Instagram Service]; Matt Buchanan, 'Instagram and the Impulse to Capture Every Moment', *The New Yorker* (online), 20 June 2013 <http://www.newyorker.com/tech/elements/instagram-and-the-impulse-to-capture-every-moment>.

88 Instagram Help Centre, *How Do I Use Hashtags?* <https://help.instagram.com/351460621611097>. Users can also tag people (other users), add their location and cross-post to Facebook, Twitter and Tumblr.

89 Tim Highfield and Tama Leaver, 'A Methodology for Mapping Instagram Hashtags' (2015) 20(1) *First Monday* [Hashtags] <http://firstmonday.org/article/view/5563/4195>.

90 Instagram Help Centre, *How Do I Use Hashtags?*, above n 88.

'likes', bookmarking or sending a post to others. The majority of Instagram activity occurs on its app because of the limited functionality of the platform's website.[91]

Before outlining what we know about systems for moderating content on Instagram, it is important to distinguish content that is removed, or self-censored, by a user from content that is removed as a result of direct intervention by the platform. Users may choose to remove content from their profiles for any number of diverse reasons that relate to individual social media practices. For instance, some users might simply want to curate their profile or reduce their total number of posts. Others, including those with larger social networks,[92] might remove images that do not appeal to the majority of their followers and/or align with their preferred self-image as part of a 'rebranding' strategy.[93] Some users might choose to remove their content due to the risk of backlash from other users, actual backlash in comments to a particular post or concerns about privacy.[94] It is also possible that users may have 'archived' rather than self-censored their content. Instagram's archive feature enables users to hide posts from their profile, including corresponding likes and comments, which a user might want to store in a private collection for their own reference.[95] To the extent that some users are choosing to remove images themselves, there is less cause for concern from the perspective of the values of the rule of law. Instagram may still have a role – and some social responsibility – in supporting or reinforcing cultural norms that impact on women's self-expression, but as previously explained, these more substantive concerns are beyond the scope of this article.

Instagram moderates content to manage the risks – financial, legal or reputational – and enormous aggregate value of users taking or uploading around 95 million photos and videos per day.[96] As noted above, the foremost way that the platform manages these risks is by reserving the unilateral right to moderate content in its Terms of Use, which stipulates that users must also comply with its Community Guidelines and other policies.[97] Instagram expressly allows and prohibits certain content in these polices, in a set of specific rules and exceptions that have changed

---

91 Alice E Marwick, 'Instafame: Luxury Selfies in the Attention Economy' (2015) 27 *Public Culture* 137, 142.

92 See generally Sauvik Das and Adam Kramer, 'Self-Censorship on Facebook' (Paper presented at Seventh International AAAI Conference on Weblogs and Social Media, MIT Media Lab and Microsoft in Cambridge, Massachusetts, USA, 8–11 July 2013) 126.

93 See, eg, Alice E Marwick and danah boyd, 'I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience' (2010) 13 *New Media & Society* 114.

94 See generally Das and Kramer, above n 92, 120–1.

95 Instagram – Help Centre, *How Do I Archive a Post I've Shared?* <https://help.instagram.com/136706673552668>.

96 Dave Lee, 'Instagram Users Top 500 Million', *BBC* (online), 21 June 2016 <https://www.bbc.com/news/technology-36584511>; Instagram, *A New Look for Instagram* (11 May 2016) <https://instagram-press.com/blog/2016/05/11/a-new-look-for-instagram/>.

97 Instagram Help Centre, *Terms of Use*, above n 23, [Content Removal and Disabling or Terminating Your Account]. Other policies include Instagram's Platform Policy: see Instagram, *Platform Policy* <https://www.instagram.com/about/legal/terms/api/>.

over time in response to the concerns of users and external stakeholders.[98] For example, the platform expressly allows photos that depict women actively breastfeeding and post-mastectomy scars.[99] This is expressed as an exception to the general prohibition on nudity, which includes, inter alia, 'some photos of female nipples' as well as 'photos, videos, and some digitally-created content that show sexual intercourse, genitals, and close-ups of fully-nude buttocks'.[100] The platform's Terms of Use and Community Guidelines are short and do not define or enumerate all types of allowed and prohibited content. Instagram's policies also contain open-textured, or redefinable, terms like 'nudity' that have, to use Hart's expression, a wide 'penumbra of uncertainty'.[101] Users are also unable to participate in Instagram's internal governance processes to learn and understand 'appropriate-versus-inappropriate distinctions'.[102] These features create a vague system of regulation for users and enable Instagram to avoid specific legal or financial obligations while also suggesting – to users, shareholders and other stakeholders – that the platform's system of regulation is organised.[103] Though the platform's outward-facing policies are vague, in practice platforms frequently review policies around content, and internal policy teams develop fine-grained guidebooks for content moderators.[104] For instance, Facebook's team of policy advisers allegedly gather every two weeks for what a senior employee calls a 'mini legislative session'.[105] Instagram is similar to other platforms in that it develops policies on an ad hoc basis in response to any number of business and other pressures.[106]

---

98 See generally Gillespie, above n 3. Instagram revised its Terms of Use and other policies in April 2018: see, eg, Instagram Help Centre, *Terms of Use*, above n 23; Brian X Chen, 'Getting a Flood of G.D.P.R.-Related Privacy Policy Updates? Read Them', *The New York Times* (online), 23 May 2018 <https://www.nytimes.com/2018/05/23/technology/personaltech/what-you-should-look-for-europe-data-law.html>.

99 Instagram Help Centre, *Community Guidelines*, above n 23, [The Long, Post Photos and Videos that Are Appropriate for a Diverse Audience].

100 Ibid.

101 Hart, above n 75, 607.

102 Magdalena Olszanowski, 'Feminist Self-Imaging and Instagram: Tactics of Circumventing Sensorship' (2014) 21 *Visual Communication Quarterly* 83, 84.

103 Crawford and Gillespie, above n 45, 418.

104 See, eg, Alexis C Madrigal, 'Inside Facebook's Fast-Growing Content-Moderation Effort', *The Atlantic* (online), 7 February 2018 <https://www.theatlantic.com/technology/archive/2018/02/what-facebook-told-insiders-about-how-it-moderates-posts/552632/>.

105 Ibid; see also Jason Koebler and Joseph Cox, 'The Impossible Job: Inside Facebook's Struggle to Moderate Two Billion People', *Vice* (online), 24 August 2018 <https://www.vice.com/en_au/article/xwk9zd/how-facebook-content-moderation-works>.

106 Crawford and Gillespie, above n 45, 419; Miranda, above n 2, 2; but see Alex Feerst, 'Your Speech, Their Rules: Meet the People Who Guard the Internet', *Medium* (online) 27 February 2019 <https://onezero.medium.com/your-speech-their-rules-meet-the-people-who-guard-the-internet-ab58fe6b9231>.

While Instagram's operators are the final decision-makers around content, a number of unknown regulatory actors – humans, artificial intelligence systems and/or other entities – undertake the work of moderating content.[107] Most of this work is undertaken by outsourced individuals or firms, known as commercial content moderators, in environments reminiscent of a call centre.[108] Recent reports suggest that 'tens of thousands of people'[109] work as commercial content moderators, including around 15 000 people who review content for Facebook and Instagram worldwide.[110] Individual moderators purportedly review hundreds,[111] sometimes thousands,[112] of posts per day. This workforce is distinct from full-time platform employees who 'are overwhelmingly white, overwhelmingly male, overwhelmingly educated, overwhelmingly liberal or libertarian, and overwhelmingly technological in skill and worldview'.[113] Moderators ostensibly make decisions about what content is acceptable primarily based on platform-specific terms, guidelines and policies.[114] Decisions around content are also influenced by the marketplace, including commercial prerogatives and responsibility to shareholders; norms at the geographic, industry, platform, community and individual levels; and state enacted laws, which platforms remain subject to, including criminal and intellectual property laws.[115] Platforms are also developing artificial intelligence tools,[116] including algorithms, to manage the enormous scale of user-generated content. The extent of influence of algorithms, however, is presently unknown.[117] This is concerning because the

107 Crawford and Gillespie, above n 45, 413 ff.

108 Sarah T Roberts, 'Digital Refuse: Canadian Garbage, Commercial Content Moderation and the Global Circulation of Social Media's Waste' (Media Studies Publications Paper No 14, Faculty of Information and Media Studies, University, 2016) 1; Roberts, 'Content Moderation', above n 2, 1; Casey Newton, 'The Trauma Floor: The Secret Lives of Facebook Moderators in America', *The Verge* (online), 25 February 2019 <https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona>.

109 Lauren Weber and Deepa Seetharaman, 'The Worst Job in Technology: Staring at Human Depravity to Keep It Off Facebook', *The Wall Street Journal* (online), 27 December 2017 <https://www.wsj.com/articles/the-worst-job-in-technology-staring-at-human-depravity-to-keep-it-off-facebook-1514398398>.

110 Newton, above n 108; Ryan Broderick, 'The Comment Moderator Is the Most Important Job in the World Right Now', *BuzzFeed News* (online), 4 March 2019 <https://www.buzzfeednews.com/article/ryanhatesthis/the-comment-moderator-is-the-most-important-job-in-the?utm_campaign=The%20Interface&utm_medium=email&utm_source=Revue%20newsletter>.

111 Ibid.

112 Weber and Seetharaman, above n 109.

113 Gillespie, above n 3, 12.

114 See generally Monika Bickert, 'Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process', *Facebook Newsroom* (online), 24 April 2018 <https://newsroom.fb.com/news/2018/04/comprehensive-community-standards/>.

115 Klonick, above n 3, 1604 ff; Crawford and Gillespie, above n 45, 412.

116 See generally Facebook Research, *Facebook AI Research* (2019) <https://research.fb.com/category/facebook-ai-research/>; see Nicolas Suzor, *Lawless: The Secret Rules that Govern Our Digital Lives* (Cambridge University Press, forthcoming 2019, copy on file with author) 109 ff <https://eprints.qut.edu.au/123199/>.

117 Weber and Seetharaman, above n 109.

technical architectures of online platforms and the algorithms designed to regulate them are value-laden and largely replicate the neoliberal and technologically determinist ideologies of platform owners.[118] Additionally, software does not always reveal why or how an algorithm reaches a particular decision.[119] Both algorithms and humans can make regulatory decisions that are potentially erroneous, biased or unfair.[120]

Users also play an important role in moderating content as an enrolled 'volunteer corps of regulators'.[121] Instagram enrols users by providing in-built reporting features, like that depicted in Figure 1, through which users can 'report' content as either 'spam' or 'inappropriate'. The platform also enables non-users to report potential violations via its webpage.[122] Klonick explains that there are many levels of content moderation:

> It [content moderation] can happen before content is actually published on the site, as with ex ante moderation, or after content is published, as with ex post moderation. These methods can be either *reactive*, in which moderators passively assess content and update software only after others bring the content to their attention, or *proactive*, in which teams of moderators actively seek out published content for removal. Additionally, these decisions can be *automatically* made by software or *manually* made by humans.[123]

Most commercial content moderation is ex post and reactive.[124] The systems of major platforms generally involve moderators allowing, denying or escalating reports of flagged content.[125] Specific policy teams usually deal with escalated reports,[126] which are often the result of policy grey areas, while artificial intelligence tools are

---

118 See, eg, Koen Leurs, 'Feminist Data Studies: Using Digital Methods for Ethical, Reflexive and Situated Socio-cultural Research' (2017) 115 *Feminist Review* 130.

119 Grimmelmann, above n 41, 1723.

120 Christian Sandvig, 'The Social Industry' (2015) 1(1) *Social Media + Society* 1, 1; Batya Friedman and Helen Nissenbaum, 'Bias in Computer Systems' (1996) 14 *ACM Transactions on Information Systems* 330.

121 Crawford and Gillespie, above n 45, 412; Julia Black, 'Enrolling Actors in Regulatory Systems: Examples from UK Financial Services Regulation' [2003] (Spring) *Public Law* 63, 84 ff.

122 Instagram – Help Centre, *Report Violations of Our Community Guidelines* <https://help.instagram.com/contact/383679321740945?helpref=page_content>.

123 Klonick, above n 3, 1635 (emphasis in original).

124 For instance, 'Facebook and most of the social media platforms have a users/community based regulation system: no one is scrutinizing the content before it is uploaded. Users have the possibility to report the posts they found inappropriate on the platform. The content moderator is basically handling these reports, called tickets': Burcu Gültekin Punsmann, 'Three Months in Hell: What I Learned from Three Months of Content Moderation for Facebook in Berlin', *Süddeutsche Zeitung* (online), 6 January 2018 <https://sz-magazin.sueddeutsche.de/internet/three-months-in-hell-84381>. See also Klonick, above n 3, 1635–6.

125 Crawford and Gillespie, above n 45, 413; Klonick, above n 3, 1639–41. Some Facebook moderators allegedly use software known as the Single Review Tool, or SRT, when reviewing individual content: see Newton, above n 108.

126 Koebler and Cox, above n 105.

increasingly moderating spam.[127] The enrolment of users as regulators is a strategic and practical solution to the problem of moderating vast amounts of content.[128] Facebook users, for instance, flag, or report, around one million pieces of content per day.[129] The in-built reporting feature also involves users in policing content and setting community norms, and is arguably 'a powerful rhetorical legitimation' for the outcomes of content moderation.[130] That is, Instagram can claim to be moderating in a way that takes into account the wishes of users, which is particularly useful in the wake of public backlash around removed content. The platform usually mitigates public backlash by, inter alia, restoring content and apologising for mistakes.[131]

127 Crawford and Gillespie, above n 45, 413.

128 Ibid 410.

129 Catherine Buni and Soraya Chemaly, 'The Secret Rules of the Internet', *The Verge* (online), <https://www.theverge.com/2016/4/13/11387934/internet-moderator-history-youtube-facebook-reddit-censorship-free-speech>.

130 Crawford and Gillespie, above n 45, 412.

131 See, eg, Aarti Olivia, 'Did We Violate Instagram Guidelines by Being "Too Fat" to Wear a Swimsuit?' *Huffington Post* (online), 14 June 2016 <http://www.huffingtonpost.in/aarti-olivia-dubey-/did-we-violate-instagram-_b_10434984.html>.

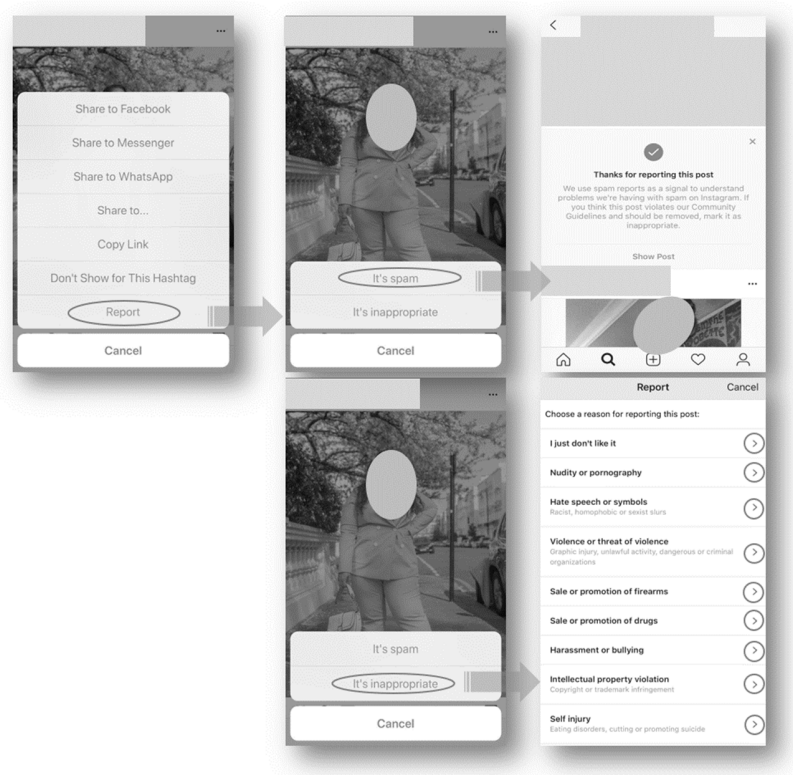

Figure 1 – Instagram's In-Built Reporting Feature for the Mobile App (iOS)

The in-built reporting feature is a sophisticated regulatory tool that can have significant implications for content moderation. On Instagram, users are able to report content as 'spam' or for any of the 'reasons for reporting' listed in Figure 1. One of the biggest problems is that the in-built reporting feature provides a very narrow 'vocabulary of complaint' that users might interpret in inconsistent ways.[132] Users can also employ the reporting function in tactical ways – as a prank, part of organised campaigns to silence particular viewpoints, or for other unclear reasons – to 'game' the system of regulation as part of horizontal conflicts between users and

132 Crawford and Gillespie, above n 45, 418.

vertical conflicts between users and Instagram.[133] While user reports do not necessarily result in the removal of content, the reporting process is problematic as there is no mechanism to verify why users report content. This means that the queue of tickets to moderate may be skewed by the biases of particular groups of users, presumably in a way that further entrenches social inequalities and other issues. Further, despite platforms largely responding to user reports, Instagram does not clearly disclose the volume or nature of actions taken to remove content from its platform.[134] There is reason to believe that the in-built reporting feature and other processes that moderate content on Instagram are far from neutral.

Powerful legal protections under US law and the operation of contract law for the most part enable Instagram, like other platforms, to resist public demands to alter its moderation processes. Section 230(c) of the *Communications Decency Act of 1996* immunises Instagram – as a 'provider' of 'interactive online services' – from liability for any content posted by users.[135] Section 230(c) has two crucial ramifications. First, platform operators, which host or republish speech, are generally not legally liable or morally responsible for what their users say or do. Platforms largely do not have to moderate content except for illegal content or content that infringes intellectual property regimes.[136] Second, if platforms elect to moderate content, they do not have to meet any particular standards of moderation.[137]

The processes for moderating content on Instagram, including the rules and guidelines that moderators follow, are also confidential information.[138] Content moderators are often required to sign non-disclosure agreements to prevent public discussion about internal decision-making processes and working conditions.[139] Some platform operators argue that the possibility of internal leaks, such as the Guardian's Facebook Files, and users gaming the processes of content moderation justify secrecy.[140] The result is that the rules and processes that moderators follow in

133 See, eg, Jean Burgess and Ariadna Matamoros-Fernández, 'Mapping Sociocultural Controversies across Digital Media Platforms: One Week of #gamergate on Twitter, YouTube and Tumblr' (2016) 2 *Communication Research and Practice* 79, 81.

134 See, eg, Ranking Digital Rights, above n 1, 60, 87–8.

135 *Communications Decency Act of 1996*, 47 USC § 230(c) (1996); Milton L Mueller, 'Hyper-transparency and Social Control: Social Media as Magnets for Regulation' (2015) 39 *Telecommunications Policy* 804, 805.

136 Tushnet, above n 45, 1001–2.

137 Ibid 1002; Klonick, above n 3, 1606–7.

138 Nyuk Yin Nahan, 'The Duty of Confidence Revisited: The Protection of Confidential Information' (2015) 39(2) *University of Western Australia Law Review* 270, 273–4 <http://www8.austlii.edu.au/au/journals/UWALawRw/2015/28.pdf>; Frank Pasquale, 'Restoring Transparency to Automated Authority' (2011) 9 *Journal on Telecommunications and High Technology Law* 235, 244–5.

139 Roberts, 'Digital Detritus: "Error" and the Logic of Opacity in Social Media Content Moderation', above n 4.

140 *Facebook Files* (2017) <https://www.theguardian.com/news/series/facebook-files>; see also Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Counterfactual Explanations without Opening the Black Box: Automated

practice significantly differ from publicly available terms and guidelines around content.[141] Controversies around images that depict women's bodies in section A below, highlight the growing unease around the unknown processes that shape the types of content that users see on Instagram.

A The Moderation of Images that Depict Women's Bodies

The moderation of user-generated images that depict women's bodies is a highly controversial issue.[142] We focus on Instagram in particular given the number of publications in news media, some of which are cited below, that make different and often conflicting claims around how the platform moderates images of female forms in practice. On the one hand, publications sometimes claim that the platform is arbitrarily removing images that depict plus-size women, stretchmarks, cellulite and a plethora of other subject matters around women's bodies.[143] Reports of Instagram censoring images of plus-size women have partially fuelled suggestions that portrayals of the Western ideal of thinness – generally 'thin-yet-toned' #fitspiration images and thin ideal #thinspiration images – are less likely to be removed from the platform.[144] Others have accused Instagram of 'blatant fat-phobia'[145] and 'fat-sham[ing]'[146] women in ways that could potentially reinforce heteronormative body standards. Purported arbitrariness and bias around content moderation are

---

Decisions and the GDPR' (2018) 31 *Harvard Journal of Law & Technology* 841, 842–3; Ananny and Crawford, above n 81, 980.

141 Madrigal, above n 104.

142 See, eg, Sara Radin, '4 Female Artists on Fighting Censorship & Reclaiming Nudes', *Highsnobiety* (online), 8 March 2018 <https://www.highsnobiety.com/p/female-artists-nude-figures-instagram/>; Elizabeth Narins, 'Instagram Deleted This Woman's Impressive Transformation Photo for Absolutely No Reason', *Cosmopolitan* (online), 21 March 2017 <http://www.cosmopolitan.com/health-fitness/a9157096/morganlosing-deleted-instagram/>.

143 Olivia, 'Did We Violate Instagram Guidelines by Being "Too Fat" to Wear a Swimsuit?', above n 131; Lindsey Lanquist, 'Amber Rose Posted a Picture of Her Bush Online, and Instagram Took it Down', *SELF* (online), 12 June 2017 <http://www.self.com/story/amber-rose-bush>; Brabaw, above n 8; Bologna, above n 10.

144 Emily Shugerman, 'This Muslim Woman Says Instagram Took Down Her Fully Clothed Selfie', *Revelist* (online), 8 March 2017 <http://www.revelist.com/religion/instagram-removed-muslim-woman-selfie/7096/so-muse-returned-to-instagram-to-make-her-feelings-known-she-reposted-the-selfie--this-time-with-a-blistering-critique-of-instagrams-censorship/3>. See also Elise Rose Carrotte, Ivanka Prichard and Megan Su Cheng Lim, '"Fitspiration" on Social Media: A Content Analysis of Gendered Images' (2017) 19(3) *Journal of Medical Internet Research* 1, 2.

145 See, eg, Gabrielle Olya, 'Curvy Blogger Angry That Her Bikini Photo Was Taken Down by Instagram: "It's Blatant Fat-Phobia"', *People* (online), 8 June 2016 <https://people.com/bodies/curvy-bloggers-bikini-photo-taken-down-by-instagram-for-violating-guidelines/>.

146 See, eg, Kristen V Brown, 'Why Did Instagram Fat-Shame These Women in Bikinis?', *Splinter* (online), 6 January 2016 <https://splinternews.com/why-did-instagram-fat-shame-these-women-in-bikinis-1793857159>.

particularly concerning as they suggest that the platform is privileging the expression of some users while marginalising others.

On the other hand, in the midst of allegations of bias, some news publications show that thin-idealised images of women are also removed from Instagram.[147] Moreover, the platform is also supposedly democratising body standards.[148] Some users claim that body positive ('BoPo') hashtags, like #embracethesquish and #effyourbeautystandards, which encourage users to accept their bodies as they are today, are acting as a counterweight to the portrayal of the thin ideal on Instagram and in mainstream media.[149] The ability of users to control their user experience – by following different users and hashtags – is significant given that counterbalanced content, including average and plus-size media, can improve individual body satisfaction.[150] The Royal Society for Public Health, which surveyed 14–24 year olds in the United Kingdom, partly echoes this theme of empowerment with the finding that Instagram makes self-expression and self-identity better for some users.[151] These various claims, among others, continue to breed confusion around the platform's processes for moderating images that depict women's bodies in practice.

The lack of publicly available information that explains why content is removed from the Instagram platform exacerbates users' concerns about alleged bias and privilege. The platform has previously noted the removal of content with the message: 'We have removed your image because it doesn't follow our Community

147 See, eg, Cambridge, above n 14; Daily Mail Australia Reporter, '"Carmen Electra Has More Revealing Photos Than Me!" The World's Hottest Grandma Gina Stewart, 48, Vows to Fight Instagram Ban over THAT Nude Photo… and Reveals the Heartbreaking Reasons She Gets Naked on Social Media', *Daily Mail Australia* (online), 11 October 2018 <https://www.dailymail.co.uk/tvshowbiz/article-6263179/The-Worlds-Hottest-Grandma-Gina-Stewart-vows-fight-Instagram-ban.html>.

148 See, eg, Evan Ross Katz, 'How Instagram Helped Democratize Beauty', *Mic* (online), 30 August 2017 <https://mic.com/articles/184143/how-instagram-helped-democratize-beauty#.vhVfC5jel>.

149 Amy Brech, 'Reclaim Your Belly: The Body Positive Movement Taking Instagram by Storm', *Grazia* (online), 6 June 2017 <http://lifestyle.one/grazia/news-real-life/real-life/reclaim-belly-body-positive-trend-instagram-confidence-embrace-squish/>. Note that #effyourbeautystandards, which model and activist Tess Holliday (@tessholliday) founded in 2011, is one of the most prominent BoPo hashtags on Instagram with around 3.8 million posts as at May 2019. News publications generally describe Holliday as one of the leaders of the BoPo movement on Instagram: see, eg, Salam, above n 14.

150 Russell B Clayton, Jessica L Ridgway and Joshua Hendrickse, 'Is Plus Size Equal? The Positive Impact of Average and Plus-Sized Media Fashion Models on Women's Cognitive Resource Allocation, Social Comparisons, and Body Satisfaction' (2017) 84 *Communication Monographs* 406; Dave Heller, 'FSU Researchers Find Plus-Size Fashion Models Help Improve Women's Psychological Health', *Florida State University News* (online), 7 June 2017 <https://news.fsu.edu/news/health-medicine/2017/06/07/fsu-researchers-find-plus-size-fashion-models-help-improve-womens-psychological-health/>.

151 Royal Society for Public Health, '#StatusofMind: Social Media and Young People's Mental Health and Wellbeing' (Research Report, Royal Society for Public Health, 2017) 23 <https://www.rsph.org.uk/our-work/policy/social-media-and-young-people-s-mental-health-and-wellbeing.html>.

Guidelines'.[152] The problem with statements like this is that the specific aspect of the Guidelines that has been breached is not clear. A plus-size blogger, for instance, stated in 2016 that 'Instagram removed my plus-size bikini post and I still have no idea why'.[153] It is important that users understand the reasons why platform operators make certain decisions given that content is, inter alia, a powerful vehicle of self-expression and a way for users to document their lives.[154] Competing narratives around empowerment and censorship, reports of selective policy enforcement and a lack of reason giving for content moderation highlight the importance of empirically examining whether seemingly like images of women's bodies are moderated alike on Instagram.[155] Given that there will always be controversies over particular instances of moderation, empirical analyses are useful to demystify how user-generated content is moderated in practice in the context of the overall moderation system. In the next Part, we outline our method for investigating Instagram's black box processes.

IV METHOD

In this study, we evaluate the extent to which the moderation of images that depict (a) *Underweight*, (b) *Mid-Range* and (c) *Overweight* women's bodies on Instagram align with the proposed values of the rule of law.[156] We develop and apply an input/output method based on black box analytics, which empirically examines how discrete inputs into a system produce certain outputs.[157] Input in this article refers to individual images while output pertains to the outcome of content moderation (ie, whether an image is removed or not removed).[158] A black box can return four types of results: true positives, true negatives, false negatives and false positives. False positives and true negatives, which we defined in Part I, are the most relevant results for this article that examines content that is not explicitly prohibited on the Instagram platform. In light of competing claims about the moderation of seemingly like images, including allegations of arbitrariness and bias, we

152 Google, 'We Have Removed Your Image Because It Doesn't Follow Our Community Guidelines' (Image Search Results, September 2017) <https://bit.ly/2WUQe4h>.

153 Aarti Olivia, 'Instagram Removed My Plus-Size Bikini Post and I Still Have No Idea Why', *Wear Your Voice* (online), 1 June 2016 <https://wearyourvoicemag.com/body-politics/instagram-plus-size-bikini-fat-woman>.

154 Kaye, *Report of the Special Rapporteur* (11 May 2016), above n 48, 4, 6.

155 See, eg, Now To Love, 'Instagram Deletes Woman's Body Positive Photo', *Now To Love* (online), 24 March 2017 <https://www.nowtolove.co.nz/health/body/instagram-deletes-womans-body-positive-photo-31390>.

156 We have received ethics approval for this research at the Queensland University of Technology (QUT Approval 1400000861).

157 Perel and Elkin-Koren, above n 21.

158 Diakopoulos, above n 21, 404.

hypothesise that images of *Underweight* women's bodies are removed at a different rate to depictions of *Overweight* women's bodies. In statistical terms, this means that there is an association between female body type and whether or not an image is removed. Our null hypothesis is that images that depict *Underweight* women's bodies are not removed differently to depictions of *Overweight* women's bodies – that is, there is no association between female body type and whether or not an image is removed.

We extracted our dataset through the Australian Digital Observatory that develops and maintains technical infrastructure for an ongoing program of research on the governance of online platforms at Queensland University of Technology's Digital Media Research Centre. Automated tools, which use the computer programming language Python, collected images that users posted to hashtags #breastfeeding, #curvy, #effyourbeautystandards, #fatgirl, #fitboy, #fitgirl, #girl, #lesbian, #lgbt, #postpartum, #skinny, #stretchmarks, #thick and #thin ('watched hashtags'). We deliberately selected hashtags, like #curvy, which news outlets have mentioned in the context of women's bodies on Instagram.[159] We also selected hashtags, such as #postpartum and #stretchmarks, because some users claim that images depicting female stretch marks and postpartum bodies are arbitrarily moderated.[160] In addition, we selected hashtags with a high number of posts, such as #girl, which has over 350 million images and videos at the time of writing, in order to generate a broader sample.[161] It is important to note that hashtags, such as #fitboy and #breastfeeding, do not necessarily accurately describe the images with which they are associated. Users employ hashtags expressively and for a number of reasons, including to reach particular audiences, or to increase the visibility of their posts.[162]

Once we selected our hashtags, the automated tools scraped the last 20 images from watched hashtags every six hours (four times per day in total) on an ongoing basis. Then approximately one month after images were collected, the availability of each image was tested again to determine whether it had been removed. The total

---

159 See, eg, Salam, above n 14; Mary Emily O'Hara, 'Why Is Instagram Censoring So Many Hashtags Used by Women and LGBT People?', *The Daily Dot* (online), 11 May 2016 <https://www.dailydot.com/irl/instagram-list-of-banned-tags-weird/>.

160 See, eg, Bologna, above n 10; Agency, 'Instagram Deletes Pregnancy Stretch Mark Pictures for Breaching Decency Guidelines', *The Telegraph* (online), 19 April 2015 <https://www.telegraph.co.uk/news/newstopics/howaboutthat/11548164/Instagram-deletes-pregnancy-stretch-mark-pictures-for-breaching-decency-guidelines.html>.

161 Dana Kilroy, '158 Most Popular Hashtags for Instagram, Marketing and More [Updated]', *Short Stack* (online), 1 February 2018 <https://www.shortstack.com/158-most-popular-hashtags-for-instagram-marketing-and-more-2017/>.

162 Highfield and Leaver, 'A Methodology for Mapping Instagram Hashtags', above n 89, [Hashtags]; Adam Mathes, 'Folksonomies – Cooperative Classification and Communication through Shared Metadata' (Research Paper LIS590CMC, Computer Mediated Communication, Graduate School of Library and Information Science, University of Illinois Urbana-Champaign, December 2004).

dataset from watched hashtags comprises 120 866 images, including 23 943 images that were removed and 96 923 images that were not removed, with an observed rate of removal of 19.8 per cent. Unlike some other platforms, Instagram does not provide a specific reason to explain why an image is no longer available. As mentioned above, this means that our programmatic tools cannot identify whether images were removed by Instagram or a user. YouTube, for example, provides detailed reasons to visitors when a video is no longer available, noting whether the video was removed by the user or by YouTube, and sometimes includes more specific information to explain which policy the video was found to violate, or what form of legal complaint YouTube received about the video. Despite this limitation, we proceeded with this study on Instagram in order to evaluate how a black box methodology might work with partial data, and to particularise the extra information that researchers may need in the future.

We generated a sample of 9582 images for the purposes of manual coding. The dataset for manual coding comprises two subsets: 4759 images that were removed and 4823 images that were not removed (an almost 50/50 split). We created our dataset for manual coding in this way because it was important that our coded sample contained a large amount of removed content to analyse content moderation on Instagram in practice. Once we generated our sample for manual coding, we undertook content analysis,[163] described in further detail below, to classify content as depicting either *Underweight*, *Mid-Range* or *Overweight* female bodies. Our coded sample, which we manually filtered from the sample of 9582 images, contains 4944 images of women's bodies. Specifically, our coded sample comprises 3879 images that depict *Underweight* women's bodies, 524 for *Mid-Range* and 541 for *Overweight*. While there was some variation in the images in each category, they generally contained the same types of content, and none of the images in the coded dataset were expressly prohibited by Instagram. Our coded sample predominantly comprises depictions of *Underweight* women's bodies despite deliberate hashtag-based selection. The prevalence of *Underweight* depictions could be due to the current thin and toned beauty ideal that is dominant in Western societies.[164]

Once we finalised our coded sample, we were able to calculate the probability or risk of removal for each category, which we outline in section B below. We

---

163 Lisa Webley, 'Qualitative Approaches to Empirical Legal Research' in Peter Cane and Herbert M Kritzer, *The Oxford Handbook of Empirical Legal Research* (Oxford University Press, 2010) 927, 941.

164 Marika Tiggemann and Mia Zaccardo, '"Strong is the New Skinny": A Content Analysis of #fitspiration Images on Instagram' (2018) 23 *Journal of Health Psychology* 1003. It should be noted that we do not believe that the coding scheme for this study resulted in *Underweight* images dominating the coded sample. While the *Underweight* category incorporates images 1–5 of the Photographic Figure Rating Scale ('PFRS'), there were less than (<) 20 images that depict women's bodies that matched images 1–2 of the PFRS in our coded sample, which suggests that the coding scheme was not overly weighted towards *Underweight* depictions.

undertook some basic quantitative analysis to ensure that the results from our coded dataset were significant in light of our sampling strategy. We used IBM SPSS Statistics to perform a chi-square hypothesis test for statistical independence between categorical variables of female body types (*Underweight*, *Mid-Range* and *Overweight*) and content removal (Removed or Not Removed).[165] A chi-square test (at a 95 per cent confidence level where $p = 0.05$) indicated a statistically significant association between results in each category, x^2 ((2, n = 4944) = 106.016, p = .000, *phi* = .146). We can therefore conclude that our results were most likely not due to random chance.

A Coding Scheme for Women's Bodies

During manual coding, we classified images as depicting either *Underweight*, *Mid-Range* or *Overweight* women's bodies. We made this distinction so that we could examine whether these categories of female body types are moderated in different ways on the Instagram platform. To determine whether images depict *Underweight*, *Mid-Range* or *Overweight* women's bodies, we referred to the Photographic Figure Rating Scale ('PFRS'), which is a measure of the naturally occurring morphology of women.[166] The PFRS comprises ten photographs of female bodies that vary in Body Mass Index ('BMI') (BMI = body weight in kilograms/height in metres squared) on a ten point scale. Images in the PFRS represent five different BMI categories: emaciated (images 1 and 2), underweight (images 3 and 4), average (images 5 and 6), overweight (images 7 and 8) and obese (images 9 to 10).[167] We chose to adopt three categories, which use the PFRS as a guiding framework, because discussion of potential censorship of women's bodies in news publications largely refers to women as either curvy or skinny (the skinny/curvy dichotomy).[168] Given that there is a high degree of subjectivity in classifying female body types, we classified images that did not clearly fall within

---

165 Julie Pallant, *SPSS Survival Manual: A Step by Step Guide to Data Analysis Using IBM SPSS* (Allen & Unwin, 6th ed, 2016) 216.

166 It should be noted that the PFRS, which some researchers have used in studies of body dissatisfaction, arguably offers a more realistic depiction of female forms than traditional contour figure (line drawn) rating scales: Viren Swami et al, 'Initial Examination of the Validity and Reliability of the Female Photographic Figure Rating Scale for Body Image Assessment' (2008) 44 *Personality and Individual Differences* 1752, 1754.

167 Ibid 1755. The 'current-ideal discrepancy' should be noted – that is, some women perceive their body type differently to their actual body type. We acknowledge that there may be discrepancy between how the subjects of each image in the dataset may classify their body type and thematic coding by Author 1: see generally Viren Swami et al, 'Further Investigation of the Validity and Reliability of the Photographic Figure Rating Scale for Body Image Assessment' (2012) 94 *Journal of Personality Assessment* 404.

168 See, eg, Sara Murphy, 'Innocent Photo of "One Skinny Woman and One Curvy Woman" Stirs Controversy', *Yahoo* (online), 16 March 2017 <https://www.yahoo.com/style/innocent-photo-of-one-skinny-woman-and-one-curvy-woman-stirs-controversy-181623226.html>.

the *Underweight* or *Overweight* categories as *Mid-Range*. The thematic categories for manual coding in this article are:

1. *Underweight*: the woman's body appears to match images 1, 2, 3, 4 or 5 of the PFRS;
2. *Mid-Range*: the woman's body appears to match images 6 or 7 of the PFRS; and
3. *Overweight*: the woman's body appears to match images 8, 9 or 10 of the PFRS.[169]

When coding the sample of 9582 images, we chose to exclude images that were explicitly prohibited under Instagram's Terms of Use and Community Guidelines, such as pornographic content. We also excluded close-ups of women's faces and depictions of two or more women with different body types. As a result, our dataset primarily comprises 'selfies' or portraits that depict a significant portion of a woman's body. It is important to note that, while there is a greater degree of subjectivity around the *Mid-Range* category, every image in this study depicts the female form in some way that is not ostensibly prohibited on the Instagram platform, so we should expect them to be moderated alike.

We undertook inter-rater reliability testing to assess the coding procedure in this study. A power analysis with 80 per cent statistical power at the 5 per cent significance level determined that 199 images needed to be inter-rated to detect more than a 20 per cent difference in the coding between the first author (Rater 1) and a volunteer (Rater 2). Then, in order to optimise confidence in the result, Rater 2 independently coded a sample of 410 images (*Underweight* = 190, *Mid-Range* = 138 and *Overweight* = 82) based on the coding scheme developed by Rater 1. We then used Cohen's kappa calculation, which is less than or equal to one where >0.5 generally demonstrates moderate inter-rater reliability, >0.7 is good and >0.8 is excellent, to test differences in coding between Rater 1 and Rater 2.[170] The coding agreement between Rater 1 and Rater 2 is excellent with a score of 0.83 for inter-rater reliability.

B Extrapolating Findings

After manually coding the images, which automated tools then classified as either removed or not removed, we calculated the probability of removal for each category by extrapolating the coded sample to a general population (or the total dataset). As noted above, the rate of removal in our total dataset is around 19.8 per

169 Swami et al, above n 166, 1755; Swami et al, above n 167.

170 See generally Mary L McHugh, 'Interrater Reliability: The Kappa Statistic' (2012) 22 *Biochemica Medica* 276.

cent, which means that we would expect most images, or posts, to remain available on the Instagram platform. Initial results from our coded sample, which we manually filtered from 9582 images with an almost equal proportion of removed and not removed content, exaggerated the probability of removal (2587 images (52.3 per cent) were removed and 2357 images (47.7 per cent) were not removed as illustrated in Table 1). We therefore needed to ensure that our results were relatively consistent with the overall 19.8 per cent (expected) rate of removal. We extrapolated the sample by: (1) calculating the proportion of removed and not removed content in the total sample – (*Removed*: 23 943/4759 = 5.031) and (*Not Removed*: 96 923/4823 = 20.095) – and (2) multiplying removed and not removed images in each category by 5.031 and 20.095, respectively.[171] Table 1 illustrates the probability of removal for each category (extrapolated to a general population) where the removed images are potentially false positives and images that were not removed are potentially true negatives.

C Limitations

The findings discussed in the following Part should be interpreted with some limitations in mind. First, like all data studies, our dataset, which is small in terms of big data, is not naturally occurring. Our sampling method, including hashtag selection, was deliberate. We limited programmatic data collection to two discrete points in time and did not capture images that users posted to private accounts. This means that only a portion of publicly available images that depict women's bodies are captured in this study. Second, as previously mentioned, our automated tools cannot determine whether Instagram or a user removed an image without knowledge of the platform's internal processes. The result is that we cannot provide data for, or comment on the proportion of content that was removed by the platform compared to content that was removed by users. We also cannot identify the precise reason why content was removed from the platform. Instagram can remove images for a number of reasons that do not directly relate to the depiction of women's bodies, such as copyright infringement,[172] and users can self-censor their content by deleting or archiving posts. This study does not focus on other factors such as race, age, disability or religion, which may also influence the moderation of women's images.

These limitations mean that this article cannot be used to make generalised findings about all images of women's bodies across all hashtags on Instagram. The data is, however, very useful for providing some insight into how images that depict

171 Note that the total dataset from watched hashtags comprises 120 866 images (23 943 were removed and 96 923 were not removed). As noted above, we generated a sample of 9582 images (4759 were removed and 4823 were not removed) for the purposes of manual coding from the total dataset. We manually filtered the sample for manual coding into our final, coded sample of 4944 images of women's bodies (2587 were removed and 2357 were not removed where 3879 are *Underweight*, 524 are *Mid-Range* and 541 are *Overweight*).

172 *Digital Millennium Copyright Act of 1998* 17 USC §512 (2000); *Copyright Act 1968* (Cth) s 116AG(1).

women's bodies are moderated on Instagram in practice. This article also makes significant inroads in developing and highlighting the importance of developing methods to probe the black box of content moderation, and we conclude with recommendations for greater transparency that will be useful to enable greater specificity in future studies of this type.

V RESULTS AND DISCUSSION

Overall, we identify a trend of inconsistent moderation across like thematic categories of images that depict women's bodies – that is, some like content was not moderated alike. The probability of removal for images that depict *Underweight* women's bodies is 24.1 per cent, which exceeds the 19.8 per cent rate of removal for the total dataset, followed by 16.9 per cent for *Mid-Range* and 11.4 per cent for *Overweight* women's bodies. Across these categories, we find that up to 22 per cent of images that were removed by Instagram or by the user are potentially false positives as depicted in Table 1. This trend supports our hypothesis that images that depict *Underweight* women's bodies are removed at different rates to depictions of *Overweight* women's bodies on Instagram.

Table 1 – Removed and Not Removed Images by Category as Percentage of the Coded Dataset (Observed) and Probability or Risk of Removal as Percentage of Extrapolated General Population (Expected)

	Underweight		**Mid-Range**		**Overweight**		**Total**	
	Observed	*Expected*	Observed	*Expected*	Observed	*Expected*	Observed	*Expected*
Removed/ Potentially False Positives	55.9% 2169 images	24.1% *10 912 images*	44.8% 235 images	16.9% *1182 images*	33.8% 183 images	11.4% *921 images*	52.3% 2587 images	22% *13 015 images*
Not Removed/ Potentially True Negatives	44.1% 1710 images	75.9% *34 364 images*	55.2% 289 images	83.1% *5808 images*	66.2% 358 images	88.6% *7194 images*	47.7% 2357 images	78% *47 366 images*
	100.0% 3879 images	100.0% *45 726 images*	100.0% 524 images	100.0% *6990 images*	100.0% 541 images	100.0% *8115 images*	100.0% 4944 images	100.0% *60 381 images*

We performed logistic regression in IBM SPSS Statistics to assess the likelihood of removal for images that depict *Underweight*, *Mid-Range* and *Overweight* women's bodies, respectively. We used the *Overweight* category as the reference group to make the odds ratios easier to interpret. As shown in Table 2, surprisingly, the odds of removal for an image that depicts an *Underweight* woman's body is 2.48 times higher than for an *Overweight* woman's body. Further, the odds of removal for an image that depicts a *Mid-Range* woman's body is 1.59 times higher than for an *Overweight* woman's body. These results suggest that there is certainly support for concerns that some like images that depict women's bodies are not moderated alike on Instagram.

Table 2 – Logistic Regression Predicting Likelihood and Content Removal for Thematic Categories (with Overweight as the Reference Group)

								95% C.I.for EXP(B)	
		B	S.E.	Wald	df	(p)	Odds Ratio	Lower	Upper
Step 1a	weight			102.345	2	.000			
	weight(1)	.909	.096	88.778	1	.000	2.481	2.054	2.998
	weight(2)	.464	.126	13.490	1	.000	1.591	1.242	2.038
	Constant	-.671	.091	54.531	1	.000	.511		

a. Variable(s) entered on step 1: weight

To the extent that content may have been removed as a result of direct intervention by Instagram, there are a number of possible explanations and contributing factors that could explain the inconsistent trend that we observe. The first is that moderators – human, artificial intelligence systems or both – could be following different sets of rules to those articulated in publicly available policies. The publicly accessible rules and guidelines of platforms are not the same as the very specific flowcharts and training materials that are used for moderation. Content moderators need to be able to quickly make consistent decisions, and operationalising the rules in practice requires them to be translated into a highly specific set of instructions.[173] It is also possible that, in response to long standing concerns that Instagram perpetuates harmful stereotypes of the thin ideal, the platform may have developed practices that are especially protective of body positivity. Moreover, while the rules that moderators follow are allegedly prescriptive, some moderators and policy teams exercise varying degrees of

173 Madrigal, above n 104.

discretion. This means that some decisions about what content is prohibited, such as 'nudity',[174] and what is not could have been made on a case-by-case basis.

Alternatively, it is possible that images in this case study were removed as a result of moderator error or bias. As previously noted, human and algorithmic moderators largely respond to reports through sorting procedures (approve, escalate or deny),[175] all within the value laden structure of the platform.[176] Moderation tasks are prioritised into queues that are influenced to some extent by algorithmic rules, which encode particular values into moderation workflows. There is also a risk that human moderators, both internal and external to Instagram, interpret and/or apply rules inconsistently based on their own value systems, or are guided by their life experience, among other things.[177] While platforms purportedly employ moderators for their language and subject matter expertise,[178] and review the consistency and accuracy of individual moderators,[179] applying standards of appropriateness to content from all corners of the globe is still an immensely complex task for moderators and platforms more broadly.[180] The potential risk of moderator error or bias is arguably exacerbated by the poor working conditions of often low-paid moderators or 'click workers'.[181] For instance, employers – whether platforms or outsourced companies – often require moderators to make decisions about content in 'a matter of seconds',[182] contributing to reliance on instinctive and value-laden responses, which are potentially problematic. We can usefully compare Instagram with its parent company here: Facebook has recently released detailed information on how its community guidelines are interpreted in practice, including providing examples of content that violates and does not violate the major terms of its policies.[183] Without greater transparency regarding how Instagram sets, maintains

---

174 Instagram Help Centre, *Community Guidelines,* above n 23, [The Long, Post Photos and Videos that Are Appropriate for a Diverse Audience].

175 Gillespie, above n 3, 74.

176 Crawford and Gillespie, above n 45, 413.

177 Sarah T Roberts, 'Social Media's Silent Filter', *The Atlantic* (online), 8 March 2017 <https://www.theatlantic.com/technology/archive/2017/03/commercial-content-moderation/518796/>.

178 See Santa Clara University, 'Overview of Each Company's Operations' (Recording of Panel Discussion, 2 February 2018) <https://santaclarauniversity.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=6e2bf22d-52cd-4e3f-9324-a8810187bad7>.

179 Koebler and Cox, above n 105. See discussion of accuracy targets and scores in Newton, above n 108.

180 Andrew Arsht and Daniel Etcovitch, 'The Human Cost of Online Content Moderation', *Jolt Digest* (online), 2 March 2018 <https://jolt.law.harvard.edu/digest/the-human-cost-of-online-content-moderation>; Gillespie, above n 3, 74 ff.

181 Buni and Chemaly, above n 129.

182 Roberts, 'Social Media's Silent Filter', above n 177.

183 Bickert, above n 114. See also Erin Egan and Ashlie Beringer, 'We're Making Our Terms and Data Policy Clearer, without New Rights to Use Your Data on Facebook', *Facebook Newsroom* (online), 4 April 2018 <https://newsroom.fb.com/news/2018/04/terms-and-data-policy/>; Guy Rosen, 'Facebook Publishes

and enforces rules around content, and monitors the performance of its moderation teams for potential bias, it is difficult to trust that content is moderated in a consistent and predictable manner.

Instagram's heavy reliance on user reporting could also explain some of the discrepancies that we identify. While user reports do not automatically result in the removal of content,[184] and the number of times users report individual content is supposedly irrelevant,[185] reporting options may still impact the outcomes of moderation. User reports are complex 'sociotechnical mechanism[s]'[186] that can be a site for political, cultural, social and other conflicts among users about the appropriateness of certain content.[187] Moderators and policy teams are tasked with arbitrating these conflicts, some of which might reflect systemic social biases.[188] However, we know very little about the inner workings of user reporting. Instagram does not report on the volume or nature of flagged and removed content, or the extent to which some content might be removed by artificial intelligence tools.[189] It is also difficult to glean information from the basic framework for user-initiated reports in Figure 1, which does not provide guidance to users to help them report content in a way that is, inter alia, consistent with the platform's policies. As a result of this lack of transparency, it is not yet possible to measure the potential impact of bias in user-initiated reports.[190]

Questions of censorship also arise in this study. As previously noted, some users suggest that Instagram's executives and/or moderators censor 'objectionable' images of women's bodies according to undisclosed normative values, and reports show a number of examples of potentially 'censorious content moderation'[191] on the platform.[192] Instagram also appears to routinely censor entire hashtags.[193] We cannot comment on the potential for direct censorship in the results that we observe given that Instagram does not report who removed content – a user or the platform – or the

---

Enforcement Numbers for First Time', *Facebook Newsroom* (online), 15 May 2018 <https://newsroom.fb.com/news/2018/05/enforcement-numbers/>.

184 Crawford and Gillespie, above n 45, 419.

185 Instagram Help Centre, *Does the Number of Times Something Gets Reported Determine whether or Not It's Removed?* <https://help.instagram.com/215140222006271>.

186 Crawford and Gillespie, above n 45, 410.

187 Gillespie, above n 3, 7 ff.

188 Renée Marlin-Bennett and E Nicole Thornton, 'Governance within Social Media Websites: Ruling New Frontiers' (2012) 36 *Telecommunications Policy* 493, 493; Crawford and Gillespie, above n 45, 413.

189 Platforms are already using some artificial intelligence tools, such as facial recognition, to match content to databases of objectionable content: see Roberts, 'Social Media's Silent Filter', above n 177.

190 See, eg, Ranking Digital Rights, above n 1.

191 See, eg, Anderson et al, above n 1, 9.

192 Gebhart, above n 54.

193 Nicolas Suzor, 'What Proportion of Social Media Posts Get Moderated, and Why?', *Medium* (online), 9 May 2018 <https://digitalsocialcontract.net/what-proportion-of-social-media-posts-get-moderated-and-why-db54bf8b2d4a>.

reasons for content removal. However, it is important to note that there are a number of actors who would ask platforms to exercise power over content for a variety of ends,[194] among them governments[195] who might seek to surveil or censor speech and copyright owners,[196] users or other third parties with grievances.[197] The result is that images in this case study may have been removed, or censored, for reasons unrelated to the depiction of female forms. Once again, the lack of public knowledge about the ways that Instagram, like other platforms, controls the types of content that users see and how and when they see, is abundantly clear.

Substantively, the higher odds of removal for *Underweight* women's bodies is somewhat surprising given the prevalence of complaints in the media alleging that Instagram favours depictions of thin, female body ideals. This discrepancy could be explained by differences in the visibility of content; perhaps thin-idealised images are more visible and therefore more likely to be reported than body positive images that are posted in a way that is visible to smaller, less acrimonious groups of users. It is also possible that different cultural norms of use among various users or communities on the platform may lead to much higher rates of self-censorship for images of underweight women, or that Instagram and its users are more supportive of body-positive images than media and blogs seem to allege. There is no easy way to tell whether these factors influence content removal without gaining access to Instagram's internal workings and/or interviewing users, which we identify as important areas for future research. While the limits of Instagram's secretive moderation processes make it difficult to come to a definitive conclusion around how, why and who removes content, our results suggest that its approach to moderation, which includes vaguely articulated rules, a heavy reliance on user reporting, and a large, generally lowly paid outsourced workforce, could lead to the inconsistency we observe in content removal.

A Evaluating Rule of Law Values through Digital Methods: Ongoing Concerns and Opportunities

Overall, our empirical examination of the ways that images depicting women's bodies appear to be moderated raises concerns about Instagram's governance practices from the vantage point of the rule of law. One significant concern is that,

194 See, eg, Kyle Langvardt, 'Regulating Online Content Moderation' (2018) 106 *Georgetown Law Journal* 1353.

195 According to the Electronic Frontier Foundation, Instagram has committed to 'reasonable efforts (such as country-specific domains or relying upon user-provided location information) to limit legally ordered content restrictions to jurisdictions where the provider [Instagram] has a good-faith belief that it is legally required to restrict the content': Gebhart, above n 54, [Overview of Criteria, Limits Geographic Scope] <https://www.eff.org/who-has-your-back-2018#limits-geographic-scope>.

196 See, eg, 17 USC §512 (2000).

197 Suzor, above n 1, 3.

users, as the subjects of regulation, lack certainty and guidance about how their content is moderated in practice. Instagram's ambiguous and incomplete policies fundamentally limit the ability of users to understand and learn the bounds of acceptable content and apply platform policies to reporting features. Raz states that '[a]n ambiguous, vague, obscure, or imprecise law [eg, Instagram's terms and guidelines] is likely to mislead or confuse at least some of those who desire to be guided by it'.[198] The fact that users are unlikely to read and understand the terms of service and associated guidelines when signing up to a platform increases the likelihood of confusion,[199] and further undermines the opportunity for Instagram's regulatory rules to effectively guide individual behaviour.

Moreover, when rules are enforced, users frequently lack transparent information and reasons to explain exactly why their content was removed, or why other users' similar content was moderated differently. This lack of information about moderation decisions breeds additional confusion and leads users to develop vernacular explanations that allege many things, including bias on the part of the platform or other stakeholders.[200] Rule of law values reinforce the desirability of users being informed of the factors that influence whether their content is visible given that content is a vehicle for users' self-expression about everyday life.[201] Content is also a conduit for users' social interactions, play and, importantly, participation in public discourses of the day.[202] However, users remain largely unguided in Instagram's complex system of content moderation that has the power to silence some female and other voices while amplifying others. These deficiencies ultimately create a highly unpredictable regulatory system, which poses an ongoing risk of arbitrariness for women and users more broadly.

This risk is compounded by Instagram's 'complete discretion'[203] to govern how users participate in its network. One of the aforementioned tenets of the rule of law is that legal instruments should limit a governing authority's power to set, maintain and enforce the law.[204] While the courts set and enforce the boundaries of contract law, the law imposes few limits on Instagram's governance in practice, which enables inconsistent and unpredictable processes for moderating content to flourish.[205] Limited checks and balances are concerning on a number of fronts, including the inherent tensions between the platform's economic interests and responsibility to

198 Raz, above n 29, 214.
199 Corinne Hui Yun Tan, 'Terms of Service on Social Media Sites' (2014) 19 *Media and Arts Law Review* 195, 197.
200 Sarah Myers West, 'Censored, Suspended, Shadowbanned: User Interpretations of Content Moderation on Social Media Platforms' (2018) 20 *New Media & Society* 4366, 4377 ff.
201 Kaye, *Report of the Special Rapporteur* (11 May 2016), above n 48.
202 See, eg, Suzor, above n 1, 2.
203 Ibid 3.
204 Raz, above n 29, 211 ff.
205 Tushnet, above n 45, 988.

stakeholders, and users' self-expression. This has given rise to deep-seated concerns that decisions around content represent the normative judgments of Silicon Valley professionals who are attempting to manage these tensions rather than giving weight to well-established Western governance values.[206]

As a society, we require ongoing conversations between platforms, users, lawmakers and other stakeholders to determine how online social spaces should be constituted and held to account. We recognise that Instagram's complex regulatory system, as part of platform governance more broadly, is a practical means of managing the sheer scale and cost of moderating vast amounts of user-generated content.[207] It would not be ideal for lawmakers to hold Instagram and other platforms to the same standards as constitutional governments. More onerous governance standards would be costly for Instagram in terms of lost revenue, for users who rely on social network technology, for the global marketplace, which relies on the flow-on effects from platform innovation, and many other stakeholders.[208] Additional regulation also raises a number of practical challenges – for instance, can and should lawmakers impose limits on the autonomy of online platforms to govern users? How can lawmakers set, maintain and enforce constitutional standards in the internet environment where the application of human rights and other laws is already unclear?

The methodology developed and applied in this article is a first step in trying to find mechanisms to constrain arbitrary power with reference to Anglo-American rule of law values. It is imperative that scholars continue to work to develop methods to independently examine the operation of moderation systems at scale, especially given the lack of transparent information about how content moderation systems work at a systemic level in practice. As part of this, it is important to take into account that major social media platforms operate at such a massive scale that errors are inevitable, and there are many interrelated factors that could contribute to actual or perceived arbitrariness. We have shown here one approach that can be used and refined to develop a quantitative indication of potential arbitrariness. Specifically, this study demonstrates the potential for digital methods that can monitor the inputs and outputs of platforms' black box systems in order to evaluate their performance. This, we suggest, is an important ongoing project that can inform contributions to public conversations around how the contemporary social spaces of the digital age are constituted and governed more broadly.

---

206 See, eg, Jeffrey Rosen, 'The Delete Squad', *The New Republic* (online), 29 April 2013 <https://newrepublic.com/article/113045/free-speech-internet-silicon-valley-making-rules>.

207 Roberts, 'Content Moderation', above n 2, 2; Tushnet, above n 45, 994.

208 Crawford and Gillespie, above n 45, 411. See also Richard Barbrook and Andy Cameron, 'The Californian Ideology' (1996) 6 *Science as Culture* 44; Electronic Frontier Foundation et al, above n 20.

Instagram can take a number of steps now to reduce the risk of arbitrariness in content moderation. A useful starting point is for the platform to clarify its terms, guidelines and policies around content. This can be done by defining key terms, explaining underlying policy reasons for content policies, providing examples of the types of content that are appropriate and inappropriate and, to the extent possible, making the internal guidelines that moderators follow publicly available. Moreover, we recommend that the platform clarifies the bounds of appropriate user reporting to better ensure that content is reported in a way that aligns with the platform's policies rather than, for instance, the individual value systems of users.

Increasing transparency and reason giving, including providing publicly available information that explains why a particular piece of content has been removed, are also key areas of improvement for Instagram. As explained above, YouTube is an industry leader in this regard; it provides a digital 'tombstone'[209] that indicates whether the removal of a piece of content was initiated by the user who uploaded it, by YouTube, or on the basis of a legal demand from a third party (for example, a copyright owner or defamation plaintiff). With this level of information, the methodology we employed in this study would have been able to provide a more definitive answer to the allegations of users that Instagram's moderation processes exhibit systematic bias against certain types of content. These allegations will not be addressed by continuing to moderate in secret; if Instagram, like other platforms, wishes to appease growing concerns about its moderation processes, it must take steps to enable some degree of external verification and accountability.[210] It is also important that users are given the opportunity to participate and articulate their views on policies that can affect their expression through content, and an option to appeal Instagram's moderation decisions affecting their individual posts.

VI CONCLUSION

The results from our empirical investigation into the moderation of images that depict women's bodies on Instagram appear to be in tension with the Anglo-American ideal of the rule of law. While the processes that moderate content behind closed doors are mostly inscrutable, we made significant inroads by identifying a trend of inconsistent removal across like categories of women's bodies. We found

209 Alex Feerst, 'Implementing Transparency about Content Moderation', *Techdirt* (online), 1 February 2018 <https://www.techdirt.com/articles/20180131/22182339132/implementing-transparency-about-content-moderation.shtml>.

210 See, eg, Tarleton Gillespie, 'Facebook Can't Moderate in Secret Anymore', *Culture Digitally* (online), 23 May 2017 <http://culturedigitally.org/2017/05/facebook-cant-moderate-in-secret-any-more/>. Interestingly, Mark Zuckerberg has recently called for greater regulation around some types of internet content, among other things: 'Mark Zuckerberg Asks Governments to Help Control Internet Content', *BBC News* (online), 30 March 2019 <https://www.bbc.com/news/world-us-canada-47762091>.

that the odds of removal for an image that depicts an *Underweight* and *Mid-Range* woman's body is 2.48 and 1.59 times higher, respectively, than for an image that depicts an *Overweight* woman's body. The overall trend of inconsistent moderation supports our hypothesis that *Underweight* depictions of women bodies are removed at different rates to *Overweight* depictions in practice, although it is not currently possible to identify whether this difference arises as a result of Instagram's direct intervention or the cultural norms of use on the platform. Interestingly, this finding suggests that claims that Instagram is less likely to remove thin-idealised images of women could be overstated.

This exploratory study provides an initial empirical application of the Western ideal of the rule of law to content moderation on social media platforms, and shows how this discourse provides a useful frame through which to evaluate and enhance moderation processes. The results of this study indicate that there are differences in how similar content is moderated on Instagram. The challenges of identifying a clear explanation for these differences raises concerns about Instagram's governance more broadly as the rules, guidelines and policies around content appear to be interpreted and enforced arbitrarily. The lack of predictability that we identified, along with deficiencies in formal equality, certainty, reason giving, transparency, participation and accountability, reinforce the need for institutionalised constraints on arbitrariness in moderation processes. These findings also underscore the need for more research and ongoing discussions about the responsibilities of social media platforms to govern their networks in a way that aligns with widely accepted governance ideals. A rule of law discourse is particularly useful here as it provides a well-established language for Instagram to identify and work through user concerns and enhance its governance practices, for users to understand and learn the bounds of acceptable behaviour, and for researchers to develop methodologies, as we have done in this article. More broadly, this discourse enables all stakeholders in platform governance to find a common ground, without the burden of onerous government structures.

Most critically, we showed the promise of using new digital methods to understand content moderation at scale, and explored some of the limitations that are imposed by the lack of transparent information that Instagram provides about the outcomes of its moderation decisions. The method in this article enables researchers to identify how much we can learn from black box analytics without open and clear rules around content, and reasons for moderation decisions. Despite these methodological challenges, we believe that it is important to continue to attempt to examine public communications on digital platforms even where the information provided is incomplete. We note an ongoing risk in internet research that certain platforms are researched more extensively because they provide easier access to data,

despite the social importance of platforms with more restrictive policies.[211] The difficulty we experienced in obtaining data and drawing explicit conclusions around how content is moderated on Instagram in practice highlights the importance of ongoing research involving black box analytics. Greater transparency around platform governance is a precondition to further, fine-grained research, which is likely to be of interest to both users and regulators.

This research continued the important work of developing the application of digital methods for empirical, legal analysis of governance on social media platforms. While this study is limited to watched hashtags and data collection at two discrete points in time, we have been able to provide one of the first legal evaluations of how Instagram moderates different images of women's bodies in practice, thus contributing to the project of digital constitutionalism. It is our hope that this research will inform debates around potential arbitrariness in content moderation processes on Instagram and, more broadly, guide the development of future legal principles for more transparent and accountable systems of regulating content on diverse social media platforms.

211 See generally, Highfield and Leaver, 'Instagrammatics and Digital Methods', above n 22, 48.