Selected Papers of #AoIR2019:
The 20th Annual Conference of the
Association of Internet Researchers
Brisbane, Australia / 2-5 October 2019

A NEW BLACK BOX METHODOLOGY: THE CHALLENGES AND OPPORTUNITIES OF INTERROGATING THE MODERATION OF IMAGES DEPICTING WOMEN'S BODIES ON INSTAGRAM

Alice Witt
Faculty of Law, Digital Media Research Centre, Queensland University of Technology (QUT)

Introduction

It is difficult to trust that user-generated content is moderated on social media platforms in ways that are free from arbitrariness and bias. 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, among them platform-specific rules and emergent social norms. Decisions around the appropriateness of content, which are made by humans and/or artificial intelligence systems, are ultimately regulatory decisions as they attempt to influence or control the types of content that users see and how and when they see it (Suzor, 2018). Content is, however, moderated within a 'black box' that obscures internal governance processes from external scrutiny. This lack of transparency has far-reaching consequences (Pasquale, 2011), one of which is that users' have limited understandings of the direct interventions that platforms make around content (Gillespie, 2018). Transparency deficits also contribute to the relative dearth of empirical research into Instagram's moderation processes to date, along with the confidential nature of the rules that moderators follow and ongoing restrictions to the Instagram Application Programming Interface (Instagram, 2019).

This paper, therefore, proposes a black box methodology for empirically examining processes for moderating content when only parts of a platform's regulatory system are visible from the outside. In doing so, it evaluates the methodological, legal and ethical challenges of studying content moderation in practice. The proposed methodology is

Suggested Citation (APA): Witt, A. (2019, October 2-5). *A New Black Box Methodology: The Challenges and Opportunities of Interrogating the Moderation of Images Depicting Women's Bodies on Instagram.* Paper presented at AoIR 2019: The 20th Annual Conference of the Association of Internet Researchers. Brisbane, Australia: AoIR. Retrieved from http://spir.aoir.org.

explained through a case study into whether like images of women's bodies are moderated alike on Instagram. None of these images are explicitly prohibited under the platform's Terms of Use and Community Guidelines. This is a topical case study given widespread user concerns that Instagram is arbitrarily 'removing' – also described as 'banning,' 'censoring' and 'deleting' – depictions of female forms (OnlineCensorship, 2018). A persistent claim is that the platform is less likely to remove thin-idealised images of women (Sarah Myers West, 2015). By contrast, some users claim that Instagram is creating a positive space for the depiction of all female forms and democratising body standards (Katz, 2017). Though this subject matter has been highly controversial over several years, users and other stakeholders lack empirical evidence to ground these competing claims.

Black Box Methodology

The proposed methodology is based on black box analytics, particularly an input/output method that identifies how discrete inputs into a system produce certain outputs (Perel & Elkin-Koren, 2017). Input in this paper refers to individual images, while output pertains to the outcome of content moderation (ie, whether an image is removed or not removed). Images were programmatically collected through the Digital Media Observatory at QUT: specifically, automated tools scraped the last 20 images from selected hashtags every six hours (four times per day) on an ongoing basis. The selected, publicly available hashtags were #curvy, #effyourbeautystandards, #fatgirl, #fitgirl, #girl, #lesbian, #lgbt, #postpartum, #skinny, #stretchmarks, #thick and #thin. These hashtags were mentioned in controversies around women's bodies on Instagram and offered a high volume of fairly diverse content.

Once images were programmatically collected, content analysis was undertaken to code images as either *Underweight*, *Mid-Range* or *Overweight* based on the Photographic Figure Rating Scale (PFRS). While coding for women's bodies is subjective, the PFRS provides a realistic measure of the naturally occurring morphology of women that is arguably more rigorous than descriptive categorisation (Swami et al., 2012). A number of images were excluded during coding, including explicitly prohibited content and close-ups of women's faces. This means that the final coded dataset (a total ('T') of 4,994 images, specifically T = 3,879 for *Underweight,* T = 524 for *Mid-Range* and T = 541 for *Overweight*) primarily comprises 'selfies' or portraits that depict a significant portion of a woman's body. We should arguably expect the sampled images to be moderated alike given that none of the images in this study are explicitly prohibited.

Approximately one month after images were collected, the availability of each image was tested again to determine whether it had been removed. This provided a discrete output for every coded input. It was then possible 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 violate Instagram's policies and were removed), across *Underweight*, *Mid-Range* and *Overweight* categories.

Results, Challenges and Opportunities

Overall, the moderation of images in this paper was inconsistent. The probability of removal for the *Underweight* category is 24.1% followed by 16.9% for *Mid-Range* and 11.4% for *Overweight.* Up to 22% of images in the coded sample were removed by Instagram or by the user and are, therefore, potentially false positives. The results suggest that claims that Instagram is less likely to remove thin-idealised images could be overstated, but that concerns around the risk of arbitrariness and, indeed, ongoing distrust of the platform among users, might not be unfounded. The empirical results are statistically significant.

The results of probing the black box around Instagram's moderation processes highlight a number of complex methodological, legal and ethical challenges (Highfield & Leaver, 2016). The foremost methodological challenge is that automated tools cannot determine whether the platform or a user removed an image without knowledge of the platform's internal processes. This means that this study, like others, cannot draw definitive conclusions about why content was removed. An added complication is that 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, and users might choose to remove their content for any number of diverse reasons. There is also legal uncertainty, largely because the legality of web scraping data remains unclear, and ethical concerns around coding images into like categories that might differ in terms of race, age, disability or other factors (Leurs, 2017).

By evaluating these methodological, legal and ethical challenges, among others, we can better assess the efficacy of using black box analytics and digital methods to examine content moderation at scale. The methodology in this paper produced a range of empirical results about the moderation of some images of female forms on Instagram and shone a spotlight on a complex regulatory issue despite the lack of transparent information about platform governance more broadly. There is wide scope for researchers to develop and apply the proposed methodology across controversies and platforms. Doing so will enable researchers and other stakeholders to continue the important work of empirically examining whether users can trust that processes for moderating content are free from potential arbitrariness and bias.

References

Gillespie, Tarleton. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Connecticut, United States: Yale University Press.

Highfield, Tim and Tama Leaver. (2016). Instagrammatics and digital methods: studying visual social media, from selfies and GIFs to memes and emoji. *Communication Research and Practice*, *2*(1), 47

Instagram. (2019, February 8). Platform Policy. *About Us*. Retrieved from https://www.instagram.com/about/legal/terms/api/

Leurs, Koen. (2017). Feminist data studies: using digital methods for ethical, reflexive and situated socio-cultural research. *Feminist Review*, *115*, 130

Onlinecensorship. (2018, 1 May). A Resource Kit for Journalists, [Instagram]. *OnlineCensorship.org.* Retrieved from https://onlinecensorship.org/content/a-resource-kit-for-journalists

Pasquale, Frank. (2011). Restoring Transparency to Automated Authority. *Journal on Telecommunications and High Technology Law 9*, 235

Perel, Maayan and Niva Elkin-Koren. (2017). Black Box Tinkering: Beyond Transparency in Algorithmic Enforcement. *Florida Law Review*, *69*, 181-221.

Sarah Myers West. (2015, November 18). Facebook's Guide to Being a Lady. *OnlineCensorship.org*. Retrieved from https://www.onlinecensorship.org/en/news-and-analysis/15

Suzor, N. (2018). Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms. *Social Media + Society*, *4*(3), 1.

Swami, Viren et al (2012). Further Investigation of the Validity and Reliability of the Photographic Figure Rating Scale for Body Image Assessment. *Journal of Personality Assessment*, *94*(4), 404.

Katz, Evan Ross. (2017, August 31). How Instagram Helped Democratise Beauty. *Mic*. Retrieved from https://mic.com/articles/184143/how-instagram-helped-democratize-beauty#.vhVfC5jel