An Investigation in the (In)Visibility of Shadowbanning

Amanda Pinto, Marquette University (United States)

International Conference on Computer Ethics: Philosophical Enquiry (CEPE) 2023, Chicago, IL

Keywords: Instagram, AI algorithms, shadowban, algorithmic bias

Extended Abstract

Social media, such as Facebook and Instagram, rely on curation algorithms to present posts for user's feeds and content moderation algorithms to prevent inappropriate, violent content from being spread. These algorithms work to create both visibility and invisibility of bodies based on trained policies of acceptability and appropriateness. Such as Instagram's policy on partial nudity, where the algorithm has been found to favor partial nudity, while community guidelines place a threshold on what is "too much" nudity. Within the liminal area of what is appropriate, where posts or users do not clearly violate community guidelines, is the technique of Shadowbanning. Shadowbanning as a practice by Instagram is only known by those who experience it, as Instagram continues to deny its existence within the algorithm. Yet the presence of users whose engagement becomes limited, the reach almost nonexistent, and discoverability low, all seen through Instagram's own analytics, suggest a technique of invisibility that presents an illusion of visibility within the algorithm. Within this paper, I will explore the technique of Shadowbanning as an interaction between the curation and content moderation algorithms through an exploration of recreational pole dancers and their proximity to what is deemed as "inappropriate" nudity.

The curation algorithm works to promote certain posts as visible to both followers and within Discovery Feeds. Taina Bucher in If…Then: Algorithmic Power and Politics demonstrates how this notion of visibility plays out within algorithms and in people's livelihoods. Bucher develops the concept of the "threat of invisibility" present in Facebook feeds, where creators are fighting to be seen for their followers. For Bucher, algorithms are important not because of what they do but because of how people shape their behavior by the algorithm. Many researchers who are critical of algorithms have shown that algorithms contain and learn many racial, sexual, gendered, classed, and able-bodied biases that influence how algorithm outputs occur. Algorithms have biases imbued within their data and reflect that in their outputs. These biases are reflected in what is made visible through curation algorithms.

In addition to curation algorithms, these social media platforms cannot exist without some form of content moderation. There are necessary parts of content moderation to keep a certain amount of violence or grotesque images from being seen. Most content moderation is done through algorithms as well, machine learning algorithms that are trained on what is acceptable and what is not, taking down content before it can be seen. In addition to the algorithm, users can "report" content and mark it for review under different categories of inappropriateness, such as violence, nudity, false information, etc. At times where posts could be flagged as more difficult for the algorithm to discern, the posts/users get turned over to human reviewers. These reviewers are trained in similar ways to an algorithm, given examples of pictures and posts of what is appropriate or not and are expected to make quick decisions on whatever posts come through their file. Meta claims that their reviewers are sort of "paired" so that cultures get an appropriate culture reviewer who can speak the same language of the post.

Where curation algorithms change the order of visibility, making certain bodies more visible than others, content moderation renders certain bodies as “inappropriate,” removing their posts and/or user profiles, making them invisible from the space. However what becomes deemed as “inappropriate” for the general public remains up to the platform to decide which becomes especially important under the category of “pornographic.” Gayle Rubin presents an extensive history of the ways that sex has been regulated and categorized into “good” or “bad” sex. She describes how “bad” sex such as sex work, pornography, or anything vaguely kink-related were regulated under obscenity laws that made owning a dildo illegal. Content moderation, through algorithms and human reviewers, continue to be hidden from public view as their training is unavailable and parameters of acceptability not shared.

Instagram’s community guidelines specifies that partial nudity is allowed but full or “excessive” nudity is deemed inappropriate for the platform and are subject to removal. This threshold become extremely tricky when one considers how plus-sized women in cheeky underwear would by nature show more skin/ butt than a thin woman’s body, leading to plus-sized women being flagged for nudity more. Importantly the guidelines for what is deemed as appropriate partial nudity versus inappropriate partial nudity were created in alignment with Victoria’s Secret guidelines for their advertisements. When does nudity and sexuality push up against the border of appriopriateness?

Curation and content moderation algorithms interact to regulate (in)visibility, presenting certain bodies and people as acceptable, while hiding bodies that are deemed inappropriate. At this intersection of curation and content moderation, on the borderline of acceptability where community guidelines are not broken lies the technique of Shadowbanning. Shadowbanning has a tricky definition, mainly because its existence is debated by communities experiencing it while Instagram continues to deny its existence. The term has arisen mainly within the groups that have felt its effects such as people of color, LGBTQ+ folks, strippers, recreational pole dancers, gender non-conforming people, and probably more. Shadowbanning does multiple things to moderate the visibility of a user and their posts: (1) hides posts from showing up in any kind of discoverability feeds and minimizes its appearance on followers timelines; (2) hides the user from being searchable unless the exact username is typed; (3) does not delete either the user or the post itself, thus giving the illusion that the algorithm is working as normal and that no community guidelines have been violated.

One community that is especially affected by this borderland of appropriateness is pole dancers. The pole dancing Instagram bubble is a large community, mostly stuck together through sharing and following of certain hashtags. Pole dancers from all over the world post and share with each other, finding community and a place where they feel comfortable posting and interacting with others. Importantly, it is also a place for pole dancers to advertise themselves as instructors and to teach other pole dancers. For many pole dancers, the ability for their content to be shared means getting more people to come to classes and thus a greater income. Similar to Instagram Influencers, pole dancers are at the mercy of the non-chronological curation algorithm that focuses on engagement. Yet they face the added technique of Shadowbanning where their visibility is managed through an unseen content moderation that renders them invisible.