Is Self-Governance Governance?: Evaluating the Consistency of Self-Governance Practices on AI in Social Media Applications

Jack West
University of Wisconsin-Madison
Madison WI, USA
jwwest@wisc.edu

Shirley Zhang
University of Wisconsin-Madison
Madison WI, USA
hzhang664@wisc.edu

Kassem Fawaz
University of Wisconsin-Madison
Madison WI, USA
kfawaz@wisc.edu

Suman Banerjee
University of Wisconsin-Madison
Madison WI, USA
suman@cs.wisc.edu

ABSTRACT

AI is prevalent in social media, powering its recommendation algorithms, face filters, and personalized advertising. Modern AI models live directly on user devices and can interact with private user data that has not been uploaded. In this work, we compare the currently deployed models for both Instagram and TikTok to their public statements about internal AI governance practices to see if these models live up to their reported standards. We found that Instagram's model may exhibit undesirable biases in concepts that their other internal models (e.g., DINOv2) may lack. We also found that TikTok's local model is less accurate in predicting age than their internal age detection system. Both findings indicate that the user should be more involved in the AI governance process within social media.

ACM Reference Format:
Jack West, Shirley Zhang, Kassem Fawaz, and Suman Banerjee. 2025. Is Self-Governance Governance?: Evaluating the Consistency of Self-Governance Practices on AI in Social Media Applications . In . ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Recently, artificial intelligence (AI) went through remarkable changes and rapidly integrated into everyday activities. Nearly 40% of U.S. adults aged 18 to 64 reported using generative AI [3]. While AI is prevalent in our everyday lives, the laws are not able to keep the same pace [28]. Currently, no federal privacy law in the US comprehensively addresses AI-driven data processing [21]. This lack of regulation leads to AI developers to *self-govern*, implementing their own organizational controls to responsibly manage AI development and deployment. For instance, Meta practices initiatives to identify and mitigate bias in their computer vision models [9, 19]. Similarly, TikTok introduces "systems" to detect underage users [23]. However, how do we know that *all* vision models made by Meta and TikTok are held to the same standards?

Mobile apps increasingly deploy AI models of multiple different modalities [22]. Initially, AI was mainly performed on powerful cloud servers [10] requiring large amounts of compute power. With the improvements to mobile devices, these powerful models are now able to run on user devices. Local models save apps time and money while offering several personalization features that were impossible with the old cloud deployment method [10]. Another benefit for good-faith researchers is **AI audits**. Local models enable researchers to assess models' fairness and safety utilizing modern security research [27].

This work explores two computer vision models found by West et al. [27] from Instagram and TikTok. Both models are deployed onto mobile devices and examine local user data. The researchers were able to directly interact with the models and inject images to evaluate their fairness across age, sex, and ethnicity. With the results from West et al. [27], we can verify Meta's claim of, "push[ing] the state of [computer vision] forward while taking steps to uncover and confront systemic injustices and help pave the way toward a more equitable future," [15]. As for TikTok, we compare the local model and their internal system on the quality of age estimation [24].

Based on our observations from Instagram and TikTok's AI model, we address the potential issues with self-governance in AI and how it relates to local data processing. Instagram's model demonstrates that users do not have any say in what labels are potentially harmful. We argue that users should define what is harmful themselves by controlling the AI models' outputs directly. We then discuss the potential challenges of giving users full control of AI processes. As for TikTok, their models are inconsistent in terms of the quality for age estimation. Their internal age estimation model is accurate, banning millions of underage users per year [24]. However, the local model deployed on user devices performs poorly for children. To address TikTok's inconsistent model quality, we discuss methods for how users can evaluate AI model quality themselves.

2 RELATED WORK

Our work explores the gap in the current cutting edge of user governance frameworks and their limitations in the context of social media and user centric designs.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

User governance frameworks for social media have previously been explored [8, 12, 17, 25]. Tyler et al. [25] performed a user study presenting participants with a novel self-content moderation framework. While the work presents interesting findings regarding ways to moderate content, it does not explore local AI moderation. DeNardis et al. [8] discuss the social ramification of commodifying human interactions. Linke et al. [12] explore governance frameworks for how to conduct online communication with social media. Both DeNardis and Linke offer data governance frameworks for the data posted by users. Our work explores a framework designed for data that is being fed to a black-box algorithm. This is novel because the user and the provider are unaware of the outcomes of said data provided to the model.

Data governance for local image AI, is more thoroughly explored in the smart city context [2, 6, 18, 20]. Smart cities consist of large scale AI computer vision systems deployed to detect real-time crime, infrastructure issues, and/or any relevant event a city may want to monitor. Choenni et al. [6] present a governance strategy where citizens whom fall victim to accidental collection of sensitive imagery can inquire for their data to be removed from the public system. The methods we explore for AI image processing, may not explicitly told to the user and are not public. Alslie et al. [2] present a distributed governance framework where a unbiased third-party AI designates when images are/aren't sensitive. Our work explores the instance where Instagram and TikTok *provide the models*. Thus, in the context of this work, the challenges we explore require that Instagram and TikTok are the AI providers.

Prior works[4, 11, 18] have discussed a user centric goverenance framework similiar to that presented in this paper. Lee et al. [11] presents a democratic framework where users instruct local models on algorithmic policy creation. Birhane et al. [4] discuss the limitations of participatory AI, which is where users work together to create and manage the best possible model. Our work differs from both Birhane and Lee as we explore the case where *only* the local model is evaluated by the user. Meaning, that each user individually crafts their own policies or corrects existing polices for themselves. Ojewale et al [18], discusses limitations in framework designs for user level audits. They claim that there are no tools that allow for the level of evaluation necessary for a full user level audit. We agree with this claim as the technology, prior to West et al., did not exist limiting capabilities of AI governance frameworks. Our work presents a new direction for user audited model frameworks.

3 MODEL EVALUATION COMPARISON

In this section, we compare the results from West et al. [27] and statements from Meta and TikTok to determine if the models abide by their own AI standards [15, 24]. We discuss how the AI governance strategies employed by Instagram and TikTok may not extend to every AI model they release. All references to model capabilities are derived from West et al.[27].

3.1 Instagram

Meta's Instagram model, denoted as `M-IG`, executes whenever an image is selected to be uploaded as a Reel. The selected image does not have to be uploaded, as `M-IG` performs the analysis on that image after selection. `M-IG` returns over **500** arbitrary concepts each given a value between zero and one. For instance, an image with an individual standing at the Washington Monument could contain the following concepts: *face: 0.985, people: 0.85, washington_monument: 0.98*. To measure fairness for these concepts generated by `M-IG`, West et al. [27] used two constructed image datasets labeled with age, sex, and ethnicity. They selected a grouping of concepts related to the face (*e.g.*, beard, blonde) and evaluated the biases, specifically, comparing sex and ethnic groups with one another. Through the two datasets, they demonstrated a potential bias across all ethnicities and sexes for all selected concepts.

Prior to West et al.'s study, Meta published their goal of strengthening computer vision model fairness [15] in 2023. They discussed two papers they published [9, 19] that aimed to make computer vision models more fair and safe. Meta first introduces FACET [9], a balanced dataset designed to benchmark demographic disparities found in computer vision models. For FACET, human annotators manually labeled images for perceived skin tone, perceived gender, and perceived age. They then evaluated their in-house model DINOv2 [19] on the FACET dataset, and it performed better than OpenCLIP [5] and SEERv2 [14], two popular and accurate computer vision models, for age and skin tones, but performed worse on gender perception. To evaluate the fairness of DINOv2, Meta first grouped 619 classes into four broad categories: Human, Possibly Human, Non-Human, or Crime. Then, they evaluated whether DINOv2 misclassified any age, skin tone, or age as non-human or associated with a crime within their dataset, which is considered harmful. DINOv2 could "classif[y] images of all groups as Human without large deviations across skin tone," [19] and showed no sign of bias against any group. With DINOv2, Meta demonstrated their understanding and commitment to computer vision fairness.

The effort that Meta contributes towards computer vision fairness is commendable. Through their work, others can follow in their footsteps and provide safe and fair models to the public. Yet, `M-IG` exhibits explicit biases across all demographics for several concepts, indicated by West et al.. We believe this inconsistency stems from how Meta's research team defined "harm." When evaluating DINOv2, Meta's human annotators determined if a concept was *harmful*. Meta's determination for what concepts were "considered harmful" was not disclosed due to the massive scope of labels. However, in the `M-IG` model, West et al. found that they considered the following labels as safe (see Table 1): location (washington_monument), religious icons (crucifix), age inferences (child and baby), and several other abstract labels (firearm, nudity, violence). We argue that the misidentification of these concepts could create new AI-related problems.

3.2 TikTok

TikTok's model, denoted as `M-TT`, is *always* active while the user interacts with the camera through the application. The model activates whenever a face is detected within the camera. Instead of arbitrary concepts captured by `M-IG`, `M-TT` measures age and sex for each detected face. The model attempts to estimate age and sex by presenting the corresponding float numbers. For example, an individual could get the following result from `M-TT` when they open up the camera in TikTok: *face_count: 1, age: 35.2205, boy_prob: 0.981322*. West et al. compared age, sex, and ethnicity biases in `M-TT`

Table 1: West et al.'s[27] analysis of the correlation of all predicted concepts of Instagram's model regarding demographic groups by running inference on our synthetic face dataset. For fair assessment, they greyed out the background in each image. They then found the following spurious correlations.

Demographic Group	Associated Concepts
Asian Man	'eyeglasses', 'bbq_barbecue', 'sansevieria', 'dais'
Asian Woman	'great_wall_of_china', 'reading', 'sports_field', 'wine', 'colHarmony'
Black Man	'rabbit', 'teamaker', 'carving', 'nighttime', 'outdoor', 'suiting', 'fish', 'chair', 'brass', 'cloud', 'balanceElements', 'RoT'
Black Woman	'video_game', 'bakken', 'drag', 'light', 'aesthetics_rating'
Indian Man	'grass', 'beard', 'skydiving', 'people', 'face', 'driving'
Indian Woman	'opening_champagne', 'confectionery', 'gamefowl', 'lepidoptera', 'jewelry', 'watchstrap', 'hair_long', 'dress', 'coffee', 'cloche', 'colVivid'
White Man	'sunglass', 'giraffe', 'businesssuit', 'water', 'indoor', 'activewear', 'sky', 'aviation', 'eyewear', 'red', 'zoo', 'nudity'
White Woman	'diningroom', 'huron', 'playing', 'sleepwear', 'lacrosse', 'blond', 'interior_design', 'fineart', 'art_painting', 'hair', 'equestrian', 'blue', 'blonde'

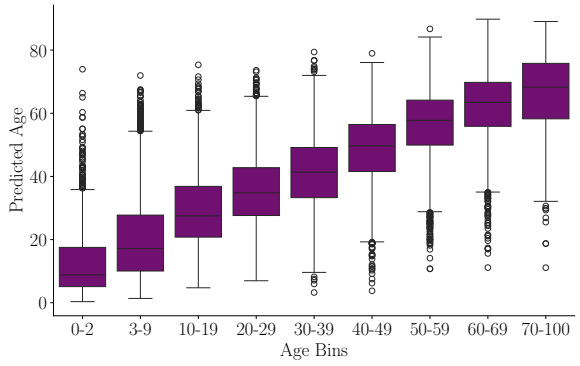

Figure 1: The distributions of TikTok's predicted age (y-axis) grouped by age bins (x-axis) are inaccurate for younger individuals as the median prediction is too high and falls outside the actual age bins. This figure originates from West et al. [27]

across each labeled value. They find that the model's age estimation is problematic: the average age estimation for toddlers and children (aged three to nine) was over 18.

The New York Times, in 2020, claimed that **over half** of TikTok's users were 14 and under [29]. To ensure children get proper guardrails, TikTok developed automated moderation technologies to find and remove content that does not align with their community standards [24]. These moderation tools extend beyond detecting underage children. They look for misinformation, AI-generated videos, and many other concepts that go against their outlined standards. Through the moderated system, TikTok banned over 60 million underage accounts in 2024 [24]. Reflecting on the reported bans, the number of underage accounts has been steadily increasing over time.

While TikTok contributes to the community by filtering out malicious or inappropriate information, and banning underage accounts, `M-TT`'s age estimation for children was not accurate (see Figure 1). It consistently classified children as adults. We observed there is inconsistency between `M-TT` and the internal models' age estimation. While TikTok can distribute and use two different models, it is paramount that AI models which interact with user data are of high quality. The fact that there is a significant gap in the model quality on user devices and the model they use privately implies a lack of consistent deployment process across different AI teams.

4 HOW CAN USERS GOVERN THEMSELVES?

As indicated in the previous section, *what companies say about AI governance does not align with what they do*. We identify two main shortcomings in current AI governance frameworks that lead to the status-quo. First, the user does not have much of a say in how AI is deployed and employed on their devices. For example, companies might employ AI locally for content moderation, where we found the Instagram model containing labels for nudity and violence. In similar settings, companies apply their governance frameworks to the development of this AI model. This means that companies determine what constitutes "harmful" content, and this determination might not include opinions from different populations. The lack of input could result in actual harm. For instance, an app providing help for people who suffer self-harm may accidentally have their content deleted by an AI which presumed the content was harmful. A similar situation has already happened when Meta removed a mental health support page for a sexually-diverse community [16] due to an AI error. Second, the AI governance policies among different teams in the same company might be inconsistent, leading to inconsistencies in policy enforcement. For example, TikTok has an online age verification model that is audited by the government. However, the local model that estimates age is highly inaccurate and appears to not be audited. It is not clear how a provider can ensure that all of their model development and deployment follow a consistent set of guidelines.

4.1 What's Next?

Both challenges reflect the *imbalanced power dynamics* and *lack of user involvement* in AI governance application. App providers can load AI models onto user devices without user involvement. We argue that if AI is going to be used locally, users should have a bigger voice in the AI governance process. One way is to offer users *full* control over local AI models. Local models offer a unique opportunity for users to fully control what data an AI model can see as all operations occur on their device. Users can directly define what they deem harmful by informing an AI model of incorrect inferences or selectively disabling concepts they do not want detected. Another way is for users to act as a distributed or crowdsourced AI model evaluator. By building tools to directly poke and prod a given AI model, users can find holes and report them.

Users take full control over AI inferences. Users can *directly* overwrite AI outputs from models or algorithms to allow for a controlled personalized experience. This control level would allow users to catch undesirable AI behavior themselves. After the AI process has occurred, all the provider would have to do is provide the AI outputs to the user in the form of a post-inference screen. The user could then view the information the AI was looking for, select the concepts and inferences they deemed undesirable, and remove them from further analysis. However, this design has several risks that we must overcome before implementation. One issue is that the output might be overwhelming to the average user. Instagram's model, for example, has too many concepts, so displaying all 500 for every picture they want to upload could lead to privacy fatigue [7]. This means that users may *not* want to be exposed to the model outputs as it could hinder their experience. Another challenge for users and providers is that some AI models perform content moderation [24]. If users were given control over content moderation models, they could circumvent measures that protect all users. Overcoming this challenge would require differentiating between content moderation and AI personalization models and not allowing users to control the content moderation models. This is a problem for providers as they must publicly standardize what is considered harmful for an AI. This standardization of harmful content already exists for profanity, Google, for example, standardized what they consider harmful language for providers to use [1].

Users evaluate model quality themselves. Different companies, or even different teams in one company, could present models that vary in quality, fairness, etc. It is up to the provider to ensure that the quality for each released AI model is up to their standards. Users can also participate in the model evaluation process as their data is what will be analyzed. We propose that, using local model processes, users can self-report model quality themselves by providing their own data as an evaluation metric. This way AI models can be locally evaluated for quality as per each user. This would also alleviate the pressure to design in-house quality assurance systems, giving developers time to further develop AI. However, social media has a large demographic of users [13] and thus some may provide poor quality evaluations. Also, due to users self-evaluating, providers will have little awareness about where the model is failing and why. To address these issues, we argue that providers should build a system that users can submit bad results to. For example, if the model failed to estimate the age of a user within an image, that user can submit that picture to TikTok to prove authenticity and help the company's models. Users who choose to provide their data directly should also be incentivized monetarily or with an in-app currency. This way, users are more likely to seek out failure cases and report them. Similar programs, like bug-bounties, have shown the benefit of crowd-sourced vulnerability locating [26].

REFERENCES

[1] [n. d.]. GitHub - coffee-and-fun/google-profanity-words: Full list of bad words and top swear words banned by Google. — github.com. https://github.com/coffee-and-fun/google-profanity-words.

[2] Joakim Aalstad Alslie, Aril Bernhard Ovesen, Tor-Arne Schmidt Nordmo, Håvard Dagenborg Johansen, Pål Halvorsen, Michael Alexander Riegler, and Dag Johansen. 2022. Áika: A Distributed Edge System for AI Inference. *Big Data and Cognitive Computing* 6, 2 (June 2022), 68. https://doi.org/10.3390/bdcc6020068 Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

[3] Alexander Bick, Adam Blandin, and David J Deming. 2024. *The rapid adoption of generative ai.* Technical Report. National Bureau of Economic Research.

[4] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the people? Opportunities and challenges for participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization.* 1–8.

[5] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2818–2829.

[6] Sunil Choenni, Mortaza S. Bargh, Tony Busker, and Niels Netten. 2022. Data governance in smart cities: Challenges and solution directions. *Journal of Smart Cities and Society* 1, 1 (Feb. 2022), 31–51. https://doi.org/10.3233/SCS-210119

[7] Hanbyul Choi, Jonghwa Park, and Yoonhyuk Jung. 2018. The role of privacy fatigue in online privacy behavior. *Computers in Human Behavior* 81 (2018), 42–51.

[8] L. DeNardis and A. M. Hackl. 2015. Internet governance *by* social media platforms. *Telecommunications Policy* 39, 9 (Oct. 2015), 761–770. https://doi.org/10.1016/j.telpol.2015.04.003

[9] Laura Gustafson, Chloe Rolland, Nikhila Ravi, Quentin Duval, Aaron Adcock, Cheng-Yang Fu, Melissa Hall, and Candace Ross. 2023. Facet: Fairness in computer vision evaluation benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 20370–20382.

[10] Kate Kaye. 2022. Why AI and machine learning are drifting away from the cloud — protocol.com. https://www.protocol.com/enterprise/ai-machine-learning-cloud-data.

[11] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, et al. 2019. WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–35.

[12] Anne Linke and Ansgar Zerfass. 2013. Social media governance: regulatory frameworks for successful online communications. *Journal of Communication Management* 17, 3 (July 2013), 270–286. https://doi.org/10.1108/JCOM-09-2011-0050 Publisher: Emerald Group Publishing Limited.

[13] Stacey McLachlan. 2024. 2024 Instagram Demographics: Top User Stats for Your Strategy. https://blog.hootsuite.com/instagram-demographics/

[14] Meta. [n. d.]. SEER: The start of a more powerful, flexible, and accessible era for computer vision — ai.meta.com. https://ai.meta.com/blog/seer-the-start-of-a-more-powerful-flexible-and-accessible-era-for-computer-vision/.

[15] Meta. 2023. Announcing the commercial relicensing and expansion of DINOv2, plus the introduction of FACET — ai.meta.com. https://ai.meta.com/blog/dinov2-facet-computer-vision-fairness-evaluation/.

[16] News.com.au. 2025. 'Will Not Be Silenced': Meta Blames 'Technical Error' for Pulling LGBTQIA+ Posts. https://www.news.com.au/lifestyle/health/mental-health/will-not-be-silenced-meta-blames-technical-error-for-pulling-lgbtqia-posts/news-story/27d7c4c58c195a90c05a27cc4523f3a7 Accessed: 2025-02-27.

[17] Jonathan A. Obar and Steve Wildman. 2015. Social media definition and the governance challenge: An introduction to the special issue. *Telecommunications Policy* 39, 9 (Oct. 2015), 745–750. https://doi.org/10.1016/j.telpol.2015.07.014

[18] Victor Ojewale, Ryan Steed, Briana Vecchione, Abeba Birhane, and Inioluwa Deborah Raji. 2024. Towards AI accountability infrastructure: Gaps and opportunities in AI audit tooling. *arXiv preprint arXiv:2402.17861* (2024).

[19] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).

[20] Krassimira Paskaleva, James Evans, Christopher Martin, Trond Linjordet, Dujuan Yang, and Andrew Karvonen. 2017. Data Governance in the Sustainable Smart City. *Informatics* 4, 4 (Dec. 2017), 41. https://doi.org/10.3390/informatics4040041 Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

[21] Privacy World. 2025. Overview of Privacy & Data Protection Laws: United States. https://www.privacyworld.blog/summary-of-data-privacy-protection-laws-in-the-united-states/

[22] Zhichuang Sun, Ruimin Sun, Long Lu, and Alan Mislove. 2021. Mind your weight (s): A large-scale study on insufficient machine learning model protection in mobile apps. In *30th USENIX Security Symposium (USENIX Security 21).* 1955–1972.

[23] TikTok. [n. d.]. tiktok.com. https://www.tiktok.com/transparency/en/combating-csea.

[24] TikTok. 2024. Community Guidelines Enforcement Report — tiktok.com. https://www.tiktok.com/transparency/en/community-guidelines-enforcement-2024-3.

[25] Tom Tyler, Matt Katsaros, Tracey Meares, and Sudhir Venkatesh. 2021. Social media governance: can social media companies motivate voluntary rule following behavior among their users? *Journal of Experimental Criminology* 17, 1 (March 2021), 109–127. https://doi.org/10.1007/s11292-019-09392-z

[26] Thomas Walshe and Andrew Simpson. 2020. An empirical study of bug bounty programs. In *2020 IEEE 2nd international workshop on intelligent bug fixing (IBF)*. IEEE, 35–44.

[27] Jackson West, Lea Thiemt, Shimaa Ahmed, Maggie Bartig, Kassem Fawaz, and Suman Banerjee. 2024. A Picture is Worth 500 Labels: A Case Study of Demographic Disparities in Local Machine Learning Models for Instagram and TikTok. In *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, IEEE, New York, NY, USA, 232–232.

[28] Esmat Zaidan and Imad Antoine Ibrahim. 2024. AI governance in a complex and rapidly changing regulatory landscape: A global perspective. *Humanities and Social Sciences Communications* 11, 1 (2024), 1–18.

[29] Raymond Zhong and Sheera Frenkel. 2020. A Third of TikTok's U.S. Users May Be 14 or Under, Raising Safety Questions (Published 2020) — nytimes.com. https://www.nytimes.com/2020/08/14/technology/tiktok-underage-users-ftc.html.