DOI:10.1145/3714417

Jack West, Jingjie Li, and Kassem Fawaz

Opinion
A Glimpse into the Pandora's Box

Demystifying on-device AI on Instagram and TikTok.

IMAGE BY ANDRIJ BORYS ASSOCIATES WITH ASSISTANCE OF SHUTTERSTOCK.AI

DEBATES SURROUNDING SOCIAL media have never stopped since its early age two decades ago. This is largely because how the social media "magic" works still remains mysterious to the public. Have you wondered what happens when you open the camera using Instagram and TikTok on your smartphone? We wondered the same circa September 2022.

Although tech companies promise that "what you do on your device will stay on the device," our prior finding reveals a leeway here: It did not stop companies from processing user data, for example, analyzing image frames on-device and sharing extracted features to their servers. For example, our initial study in 2022 found CISCO Webex was extracting audio metrics from a videoconference call while the user was muted[8]—a practice they stopped after we reported our findings.[5] Since then, there have been significant advances in artificial intelligence (AI) techniques and computation resources on smartphones. Our hypothesis was that nothing prevents an app from applying similar AI models to cellphone cameras, on which users have little awareness and control. This column describes how we investigated this hypothesis over the subsequent two years.

Before we get into details, we must answer a question: Why is it important to know what the apps do if the raw data (for example, camera frames, user images, live audio) never leaves the device? When tech companies first deployed AI models, they lived on the cloud, requiring user data to leave the device. These methods inadvertently created a clear separation between user data and AI models. As these models were part of the cloud, users had nothing to worry about if their data stayed on the device.Recently, AI has migrated to our devices, making predictions locally.[2] The migration of AI models from the cloud to user devices is natural. Local models no longer take resources from the cloud, algorithms can now be dynamic and adapt to user interests, and users' data remains on their devices. However, AI now has direct and immediate access to user data, so the clear separation that once existed is gone. Because of that, local AI models have more access to user data than ever. Before local AI processing, applications could not examine camera frames without them leaving the device. Now, unbeknown to the user, an AI model can analyze camera frames in real time, extracting sensitive features and concepts. We argue that users have the right to know what information apps extract from their data. Increased transparency can better inform social media users of model capabilities and potential risks, including identity misrepresentation.

The trend of migrating AI algorithms to devices makes it possible to understand how they process user data. In

September 2022, we started analyzing the two most popular social media apps on Android: TikTok and Instagram. We thought that apps would call APIs from popular libraries. So, we decompiled the apps to look for these APIs, but we did not find them. We tried applying some of the proposed techniques for local model extraction,[6] which did not help either. It turns out we needed to dig deeper because Instagram and TikTok employ sophisticated code obfuscation with low-level execution of AI methods. We developed new reverse-engineering techniques to determine when AI processes occur, what the inputs and outputs from the model were, and what happened to the outputs. In particular, we created our custom operating system that tracked all activities done by each application. We then used the application as usual and looked for evidence of AI processes. Once found, we performed dynamic instrumentation to interact with the models directly. We found two computer-vision AI models for both Instagram and TikTok.[7]

TikTok triggers a vision model while the camera is open, and the camera sees a face. To be specific, we found that TikTok's model feeds every single frame from the camera to their local model, extracting demographic information about the user. TikTok estimates the user's age and gender and draws a bounding box around the user's face. The outputs from the model are then written to an encrypted file. We found no evidence that the data left the device. Since the model's outputs are tied to the existence of a face within a camera frame, we showed TikTok hundreds of thousands of faces from the FairFace dataset. Each face is labeled with a gender, an age, and a racial demographic. We found that TikTok's model commonly overestimates the age of children. The model's average age estimation for babies and toddlers was 13 (TikTok's minimum age for making an account).[7] We also found that gender identification was highly inaccurate for Black individuals.[7] If these values were to be used for age verification, children would be able to easily bypass the model. No mention of this risk is expressed in their current privacy policy.[a]

Instagram, on the other hand, executes a vision model when the user selects an image to be uploaded as a Reel.[b] The user does not need to post the image to trigger the AI process: The model consistently executes upon selecting an image. The input to the model is the chosen image, and the output is over 500 different concepts. The concepts vary widely; there are landmarks (Washington Monument, Great Wall of China), facial features (beard, blonde), and objects (menorah, crucifix, ball). To evaluate risks to users, we constructed a synthetic dataset to measure the biases associated with different racial demographics. Using this custom-made dataset, we evaluated which racial demographics have higher scores than others for the various concepts. For example, we found that an Asian woman is highly correlated with the Great Wall of China, and a White woman is similarly correlated with blonde.[7] Our work demonstrated that if these models were used for algorithmic decision making, they would exhibit a significant bias. Due to the models' black-box nature, the consequences of the biases are unknown.[1] However, this lack of understanding has not impeded AI development. After publishing our work, we reanalyzed Instagram and found a new model with more than 1,500 concepts. There appears to be no slowing down for AI on local devices.

The trend of migrating AI algorithms to devices makes it possible to understand how they process user data.

How do the users feel about these models? To answer this question, we performed a systematic user study of social media users to understand how these models impacted their usage of the apps. We recruited 21 Instagram and TikTok users and interviewed them about their perceptions and understandings. We first asked about their knowledge of AI, led participants through a lesson about computer vision (so that everyone has similar understanding), and then revealed the models. Then, we asked them how they understood AI and if exposure to the AI models used by Instagram and TikTok fit into their current understanding. Two weeks later, we asked the participants if their usage of Instagram and TikTok changed.

Participants were shocked by Instagram and TikTok in different and interesting ways. They indicated TikTok's algorithm for processing images is non-consensual and invasive. Immediately using AI without any indication felt wrong to them. Participants also reflected on new fears of accidentally opening the camera and the model processing sensitive images. Their reaction to TikTok revealed a significant gap in user understanding of AI capabilities. Participants assumed they could infer AI behaviors based on the app's speed or believed that AI was tied to specific interactions they had with the apps. For example, most participants thought what they liked on TikTok was where the bulk of the AI processing took place, not their raw images. We found that several parents who participated in our study were concerned about what the app could infer about their children. One parent was offended that the app was inferring gender as their child was non-binary. All participants expressed they were unaware of this possibility and wanted more transparency from the application.

Participants were more positive toward Instagram because they liked that they could control the model's execution. Because a user had to select an image, it felt more consensual, as they could directly connect an action to the AI. This control gave participants agency when interacting with Instagram, which was noted positively. However, participants expressed negative feelings toward the amount of data gathered from their single image. Due to the sheer amount of concepts that Instagram's model produced, it was hard to understand the purpose of the data. This confusion forced participants to infer the meaning based on prior assumptions of the app. Participants were skeptical even when informed that the

a Privacy Policy (TikTok); https://bit.ly/4g79688

b Privacy Policy (Instagram); https://bit.ly/3E4E7fD

data may not leave or be used by the device at all. To them, their data was why social media was free; thus, why would Instagram make a model and not use the data?

Overall, all participants expressed that they wanted apps to be more transparent with how they used AI so they could avoid it if they desired. Out of the 21 participants we interviewed, eight reported declining usage due to our intervention. For example, parents reported they informed their children of the risks, and they disabled the camera permission. Others said they were more aware about where they use Instagram and TikTok. One participant uninstalled TikTok after our study. It appears that making users more aware of how apps process their data affects their behavior. This begs the question: How should apps provide more transparency about their internal AI processing?

How applications should inform users about their AI use is not clear. Do the users need yet another privacy notification, menu, permission, popup, or policy? Not necessarily. Still, we can leverage the recent push for privacy or data-safety labels for mobile apps.[3] These labels, modeled after privacy nutrition labels, are supposed to be a concise and simple way to summarize an app's privacy practices. At a high level, they cover types and purposes for data collection and sharing. We propose that these sections be amended with practices around local AI processing. Here, there are several challenges the research community, developers, and platform providers must come together to address.

The first challenge is what data to include in these newly created labels. They can discuss the data the model analyzes, the analysis frequency, and what triggers the analysis. Also, they can include the model outputs and purposes. Given the increasing popularity of local language models, which enable reconfigurable analytics at runtime, we envision a combination of measures, in particular compliance enforcement and model certificates, should be developed and employed in conjunction with the safety label. In addition, a model benchmark can be useful for users and developers, providing insight into model performance in different contexts. The second challenge is how to display all this information in the label format in an accessible and informative way. AI-related concepts are complex, and users might not fully understand their ramifications. As such, interfaces should prioritize user understanding in how they present AI-related information. Personalized interfaces will help ensure that even non-expert users can comprehend the nature of the processing and its implications. Finally, a critical challenge is who is responsible for creating and auditing these labels. Developers face challenges in completing privacy labels in apps,[4] and we foresee similar challenges for these AI cards. More importantly, AI and social media governance, including the deployment of the AI safety label, is a global effort. However, it is unrealistic to believe that policymakers across the world will reach their consensus sometime soon, as evident again by the recent episode of the U.S. TikTok ban and social media "refugees."[c] The unstoppable growth of AI capabilities leaves researchers with an open question: How can we inform and empower users to navigate the chaos before standardized safety measures are put in place? C

c See "Chinese app RedNote gained millions of U.S. users this week as 'TikTok refugees' joined ahead of ban"; https://bit.ly/4gc85fg

References

1. John-Mathews, J.-M. Critical empirical study on black-box explanations in AI. In *Proceedings of ICIS 2021: 42nd Intern. Conf. on Information Systems* (2021).
2. Kaye, K. Why AI and machine learning are drifting away from the cloud. *Protocol.com*; (Aug. 2022); https://bit.ly/4hBA09A
3. Kelley, P.G., Cranor, L.F., and Sadeh, N. Privacy as part of the app decision-making process. In *Proceedings of the SIGCHI Conf. on Human Factors in Computing Systems*. ACM (2013); 10.1145/2470654.2466466
4. Khandelwal, R. et al. Unpacking privacy labels: A measurement and developer perspective on google's data safety section. In *Proceedings of the 33rd USENIX Security Symp. (USENIX Security 24)* (2024).
5. Kulkarni, A. Innovation behind Webex mute. *Webex Blog* (Apr. 2022); https://bit.ly/3PHr7zj
6. Sun, Z. et al. Mind your weight (s): A large-scale study on insufficient machine learning model protection in mobile apps. In *Proceedings of the 30th USENIX Security Symp. (USENIX Security 21)* (2021).
7. West, J. et al. A picture is worth 500 labels: A case study of demographic disparities in local machine learning models for Instagram and TikTok. In *Proceedings of the 2024 IEEE Symp. on Security and Privacy (SP)*. IEEE Computer Society (2024).
8. Yang, Y. and West, J. Are you really muted?: A privacy analysis of mute buttons in video conferencing apps. In *Proceedings on Privacy Enhancing Technologies* (2022).

Jack West (jwwest@wisc.edu) is a Ph.D. student at the University of Wisconsin-Madison, Madison, WI, USA.

Jingjie Li (jingjie.li@ed.ac.uk) is an assistant professor at the University of Edinburgh, Edinburgh, Scotland.

Kassem Fawaz (kfawaz@wisc.edu) is an associate professor at the University of Wisconsin-Madison, Madison, WI, USA.