Tennessee v. Meta
Case No. 23-1364-IV
PX-00550A

# Follow-Up: Cosmetic Surgery Effects Execution Plan

## A. Goal

Following the feedback on the Cosmetic Surgery AR Effects review, this document will explain how we would approach the execution of *Option 2 (Lift the ban on Cosmetic Surgery AR Effects but remove these effects from IG recommendation surfaces)* or *Option 3 (Lift the ban on Cosmetic Surgery AR Effects)*. We're also including a description of maintaining the status quo *(Option 1)*.

## B. Recap: Decision Needed & Options

*Copied from [pre-read] Cosmetic Surgery Effects*

Should we continue, modify, or lift the ban on Cosmetic Surgery AR Effects that we instituted in October 2019 in response to expert and press concerns about potential negative wellbeing impacts (especially on minors)?

- **Option 1: Continue the temporary ban with current Policy and re-evaluate when there is more definitive data on wellbeing and product impacts [FB App and Instagram Policy + Comms, Responsible Innovation Team recommendation]**
    a. Pros: Mitigates wellbeing concerns; no PR/regulatory risk
    b. Cons: Limits growth; limits voice and innovation; arguably western-centric
- **Option 2: Lift the ban but remove these effects from IG recommendation surfaces [FB App + Instagram Recommendation / Spark AR + MSGR + i18n open to a version of this]**
    a. Pros: Fosters voice and innovation; lower impact to growth
    b. Cons: Still notable wellbeing risk (since recs are marginal source of virality); possible bad PR
- **Option 3: Lift the ban [Boz/Spark AR, MSGR + i18n recommendation]**
    a. Pros: Fosters voice and innovation; lowest impact to growth; supports regional perspectives
    b. Cons: Highest wellbeing risk; possible bad PR

**CAVEAT**: Data on product impact is limited, and research into wellbeing impacts of AR effect is too new and not yet robust. Meaningful research may take years. Therefore, we must make a **value-based** decision here -- essentially, safety vs. voice.

## C. How would we approach each option?

### C.1 Summary

- **Option 1**
    - Comms Approach: (*Minimal-to-No Risk* for Press and Policy Scrutiny) Continue reactive comms posture (e.g., sharing our interim policy language externally).

- Product: On-going human review (+ 15 weeks if we decide to build classifiers, etc. for long-term sustainability).
- Timing: n/a

- **Option 2**
  - Comms Approach: (*High Risk* for Press and Policy Scrutiny) Proactive announcement and brief vocal press and creators in the Spark community, using AR Spokespeople. Reactive statements ready to address any mistakes made in the moderation process, if one-off cosmetic effects accidentally make it onto IG recommendation surfaces.
  - Product: Will take 15 weeks to implement + on-going human review.
  - Timing: ~15 weeks when product is ready.
- **Option 3**
  - Comms Approach: (*Highest Risk* for Press and Policy Scrutiny) Proactive announcement to press and creator group. Brief top creators to serve as third party references for the press. Brief press that have covered our interim ban.
  - Product: Will take 1 week to implement.
  - Timing: ~6 weeks when COVID conditions have stabilized and near-term Facebook & Instagram COVID well-being programs have launched.

## C.2 Communications Execution Summary

**Option 1: Continue the temporary ban with current policy and re-evaluate when there is more definitive data on wellbeing and product impact.**

- **Comms/Policy Risks: *Minimal-to-No Risk.*** Press globally have already reported on the ban as a formal policy change even despite our messaging it as interim, and this is a topic where we come under regular scrutiny in relation to the negative impact on mental health of [redacted] people. We'll use this data to align with our well-being efforts and commitment to combat the impact of these effects.
- **How we'd approach:** Reactive. Continue sharing our interim policy language externally, if asked.
- **Potential Global Headlines:** N/A *(given this was previously reported on)*

**Option 2: Lift the ban but remove these effects from IG recommendation surfaces**

- **Comms/Policy Risks: *High Risk.*** External stakeholders (press and policy stakeholders) will pick up on our loosened policy and question our commitment to supporting well-being on IG, leading to negative press globally. Many creators will be relieved, although vocal critics could accuse Instagram of unfair/inconsistent policing i.e.allowing effects to be published, but not promoting them.
- **How we'd approach:**
  - Proactive announcement, aimed to address our creator community through a post in the Spark AR Facebook group, and brief handful vocal reporters who previously covered this news, using AR Spokespeople.
  - Message that we want to strike the right balance between well-being and manage the feedback we've received from the creator community that the ban has gone too far in restricting artistic license and creativity.

- Reiterate the guidelines and policies we have in place to select content for placement in recommended surfaces, leaning on Facebook's Spark AR Platform Policies and IG's community guidelines.
- Prepare reactive statements if we make a mistake in the moderation process (i.e. if one-off cosmetic effects accidentally make it onto the platform), and note that users also have the option of reporting an effect they feel is inappropriate or violates our guidelines.

- **Timing:**
  - **Best Case:** Align with product teams on timing, and introduce changes in ~15 weeks when COVID conditions have stabilized and near-term Facebook & Instagram COVID well-being programs have launched. We do not want to undermine any new product launches or risk harming the goodwill and positive sentiment we're building.
  - **Second choice:** Introduce changes when COVID has stabilized in ~6-8 weeks, if this timing aligns when the product team is ready. Earlier timing is less than ideal as we'll introduce during a period of new product launches aimed at well-being and supporting our user's mental health during this vulnerable time. At best, positive coverage from other programs could drown this out. At worst, press could see it as confusing and duplicitous.
- **Anticipated Global Headlines:**
  - **In 15 Weeks At Launch:**
    - Instagram allows harmful cosmetic surgery effects back onto the platform
    - Instagram reverses its ban on plastic surgery filters to maintain growth
    - Instagram perpetuates pressure to be perfect, despite prior wellbeing efforts
    - Facebook and Instagram are backtracking their efforts to support healthier body image
    - Instagram loosens ban on cosmetic surgery effects, reaching middle ground with Spark AR Creators *(best-case)*
    - In a win for creators, Instagram rolls back plastic surgery filter ban (*best-case*)
    - Cosmetic surgery effects are back on Instagram, but you'll have to try harder to find them
  - **If announced earlier - same headlines plus:**
    - Instagram Thinks That Now Is A Good Time For You To Be Using Cosmetic Filters
    - You may not be able to leave your house, but you can once again plump your lips
    - With a captive audience, Instagram reintroduces cosmetic surgery effects

**Option 3: Lift the ban altogether**

- **Comms/Policy Risks:** ***Highest risk.*** We expect strong criticism that we are backtracking on our current ban, that press globally have already reported on as a formal policy change even despite our messaging it as interim. This decision will jeopardize our earned credibility of being a platform that prioritizes well-being, thus putting any future related campaigns at

risk. Creators may criticize us for arbitrary decision-making that caused months of lost creative output, but will be glad to see the changes revert.

- **How we'd approach:**
  - Proactive announcement, aimed to address our creator community through a post in the Spark AR Facebook group, and brief handful vocal reporters who previously covered the interim ban, using AR Spokespeople.
  - Brief top creators to serve as 3P references for press.Our goal would be to manage negative sentiment as much as possible. In messaging, we'll lean on creativity and expression, and point to the creator community's desire for more artistic license and creativity in effect creation.
  - Reiterate the guidelines and policies we have in place to select content for placement in recommended surfaces, leaning on Facebook's Spark AR Platform Policies and IG's community guidelines.
  - Be ready with reactive statements if we make a mistake and note that users also have the option of reporting an effect they feel is inappropriate or violates our guidelines.
- **Timing:**
  - **Best Case:** Introduce changes in ~6-8 weeks when COVID conditions have stabilized and near-term Facebook & Instagram COVID well-being programs have launched. We do not want to undermine any new product launches or risk harming the goodwill and positive sentiment we're building.
  - **Second choice:** Introduce changes in ~2 weeks. Less ideal as we'll introduce during a period of new product launches aimed at well-being and supporting our user's mental health during this vulnerable time. At best, positive coverage from other programs could drown this out. At worst, press could see it as confusing and duplicitous.
- **Anticipated Global Headlines**:
  - In Six Weeks:
    - Instagram does a total reversal on cosmetic surgery effects and their spread
    - Instagram is perpetuating "Instagram Face" and jeopardizing the mental health of users, despite prior efforts to the contrary
    - Facebook and Instagram's Spark AR policies create confusion among the developer community as previously banned cosmetic effects are now allowed
    - In a win for creators, Instagram rolls back plastic surgery filter ban
  - In Two Weeks - same headlines plus:
    - Instagram caps a week of well-being launches with...cosmetic surgery filters
    - Cosmetic surgery filters are back on Instagram, just in time for the COVID-15

## C.3 Product Execution Summary

- *For all Options,* we will need to handle the 37K effects in the submissions queue (approx 9-10 days worth of total daily effect submission volume), that have been in limbo since we established the temporary ban in Oct 2019. [1 Week]
- *For Options 1 & 2:*

- We will need to build the integrity workflows or classifiers that allow us to reliably detect cosmetic surgery effects and automate this process. [10 Weeks]
- Given the complexity and subtlety required to make an accurate decision, we will need to use human review to ensure accuracy. Human review operations are variable costs and will scale as the number of effects increase. [4 Weeks]

- Not included here is any additional product work that may be explored in the future.

This table outlines the product work that would need to happen for each option. A more detailed explanation of this work is in the appendix.

| | OPTION 1<br>**Continue ban** | OPTION 2<br>**Don't recommend*** | OPTION 3<br>**Lift ban** |
|---|---|---|---|
| **EXISTING QUEUE HANDLING**<br>*EFFORT: 1 WEEK* | | | |
| **BUILD CLASSIFIERS**<br>*EFFORT: 10 WEEKS* | | | |
| **SET UP HUMAN REVIEW AT SUBMISSION**<br>*EFFORT: 4 WEEKS +*<br>*On-going investment* | As submission volumes rise we would need to continually increase the number of human reviewers until we reached high levels of accuracy with automation. | | |
| **SET UP HUMAN REVIEW FOR RECOMMENDATION**<br>*EFFORT: 4 WEEKS +*<br>*On-going investment* | | Only review content that meets recommendation criteria. This should keep the total number of human reviews lower but still not fully scalable as we automate recommendation. | |
| **TOTAL EFFORT (EST)** | 15 WEEKS + | 15 WEEKS + | 1 WEEK |

[ ] Work to be completed
[ ] No work needed

*AR recommendation surfaces only drive 1.5% of Platform Effect Stories Sharing DAP

Effort is an estimate of the time required if half of the AR Integrity team (3 ENG + XFN) were dedicated to the project.

# Appendix I - Comms Background

**Current state of affairs**

- Prior to the interim cosmetic surgery ban, policy makers, well-being experts, and media heavily criticized Instagram for its perceived negative impact on body image and mental health for [redacted] people. AR filters were specifically cited as a cause for the growing popularity and normalization of plastic surgery.
- Far from being seen as overly paternalistic, Instagram received global positive support from press, policy makers and the broader public when the temporary ban on cosmetic surgery effects - known to be heavily used by teen girls - was announced. It is worth noting that we informed media that we were re-evaluating our policies, however, media reported this as an official ban of all cosmetic effects.
- Regulators continue to inquire about the negative impact social media plays on body image issues. Just this week the UK Parliament's Women and Equalities Select Committee announced an inquiry into body image and specifically references the effect of content promoting cosmetic surgery/interventions on body image when using social media. In 2019, the UK's Digital, Culture, Media & Sport Committee asked Instagram to consider discontinuing filters that create unfair pressure on teen body image. Instagram committed to taking a harder look at the issue.
    - An internal exploratory study of social comparison and body image among teen users in the US, conducted by Instagram, showed that 52% of teen girls that experienced negative social comparison on Instagram said it was caused by images related to beauty. Among teen girls that said they felt bad about their bodies, 32% said Instagram made them feel worse.
- Reversing this decision undermines Instagram's well-being commitment and jeopardizes any future efforts to work with policymakers and mental health experts on these issues. Should we decide to lift or loosen the policy, we will message the change as tied to striking the balance between expression and safety.

# Appendix II - Product Execution Details

## Queue Handling (required for all options)

***For 37K cosmetic surgery effects in the holding queue:*** we are aware that some creators have re-submitted very similar effects without the cosmetic surgery elements, and we expect that because of the time that has passed since the temporary ban some effects are now abandoned (55% of creators with effects in this queue have not been active during March).

**Given that we would:**

1. Set all effects back to a draft state and notify creators so they can resubmit their effects where appropriate.
    - This will reduce the load on the Spark AR review process compared to adding abandoned effects to the queue, as it is already under considerable strain given reduced review capacity due to COVID-19.
    - It also prevents creators from being surprised by their content going live unexpectedly.

***For all new submissions:*** we would remove the system that holds cosmetic surgery effects and these would be placed into our existing review process.

## Human Review & Automation (options 1 & 2)

For either option 1 or 2, we would need to build out human review flows as we believe it is unlikely that we can build classifiers with a high enough level of accuracy, at the moment.

Therefore both options 1 & 2 would require a combination of automatic classification and then human review. For a complete ban (option 1) we would need to review all submissions where we see potentially violating components, for a relaxation (option 2) we would only need to review effects before they are recommended.

**Contextual information:**

- Recommendation surfaces only drive 1.5% of Platform Effect Stories Sharing DAP. The majority of our usage is from users organically discovering effects through Instagram Stories.

**Implementation for full ban (option 1):**

1. **Existing effects:**
    a. We would need to re-review effects on our platform to label whether they meet the cosmetic surgery threshold or not.
    b. This could be partially automated but could require circa 50-60 days human reviewing time, we would recommend declaring an amnesty on low impression effects and only review once past an impressions threshold, which would reduce this substantially.
2. **On submission**:

a. Use classifiers to determine effects which could contain elements that fall foul of plastic surgery policy and pass to human review before acceptance onto the platform.
b. Review all identified effects for cosmetic surgery policy:
    i. We would need to build new review flows that focus on cosmetic surgery treatments, so human reviewers can quickly and reliably notice them. This is estimated to take 4 weeks of engineering time.
    ii. We expect that the new policy would increase our average review time (Circa. 50s to 70s+).

**Implementation of relaxation of ban - no recommendation (option 2):**

1. **Review for recommendability:**
    a. All effects are default not recommendable until they are human reviewed.
    b. Use classifiers to determine effects which could contain elements that fall foul of plastic surgery policy and pass to human review when they meet recommendability criteria (e.g. high engagement, viral, top creator).
    c. Review recommendable effects and build out the same tools as used for human review above but only use if the effect could be recommended to reduce human review load.