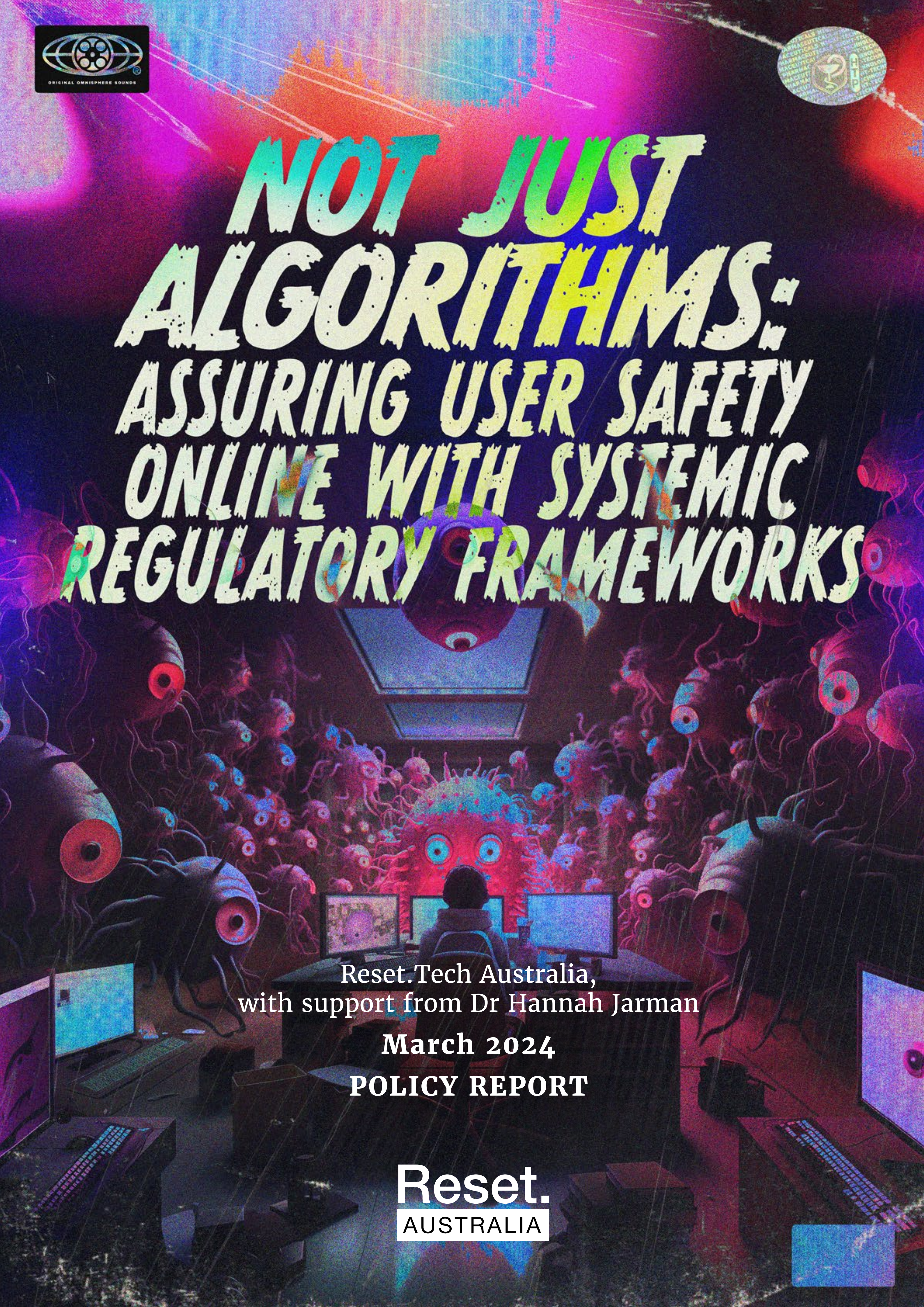
NOT JUST ALGORITHMS: ASSURING USER SAFETY ONLINE WITH SYSTEMIC REGULATORY FRAMEWORKS



Reset.Tech Australia,
with support from Dr Hannah Jarman

March 2024

POLICY REPORT

Reset.
AUSTRALIA

Trigger warning.

THIS REPORT CONTAINS IMAGES AND DEPICTIONS OF EATING DISORDERS AND SELF-HARM. ALL IMAGES INCLUDED IN THIS REPORT HAVE BEEN ACCURATELY RECREATED FROM ORIGINAL CONTENT, BUT IDENTIFYING DETAILS HAVE BEEN REMOVED.

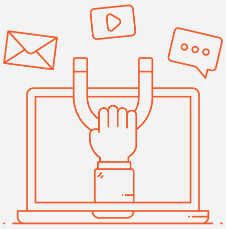
Report production by
Benjamin Horgan Design Studio

Icons used throughout report
were created by ‘Brickclay’
from Noun Project
CC-BY 4.0

SUMMARY

Many of the systems and elements that platforms build into their products create safety risks for end-users. However, only a very modest selection have been identified for regulatory scrutiny. As the government reviews the Basic Online Safety Expectations and *Online Safety Act,* the role of *all* systems and elements in creating risks need to be comprehensively addressed.

This report explores the role of four systems (recommender systems, content moderation systems, ad approval systems and ad management systems) in creating risks around eating disorders. We ran experiments on a range of platforms (including TikTok, Instagram, Facebook, X and/or Google) and found that:

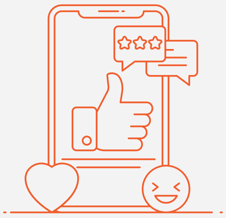


01. Content recommender systems can create risks.

We created and primed 'fake' accounts for 16-year old Australians and found that some recommender systems will promote pro-eating disorder content to children.

Specifically:

- On TikTok, **0%** of the content recommended was classified as pro-eating disorder content;
- On Instagram, **23%** of the content recommended was classified as pro-eating disorder content;
- On X, **67%** of content recommended was classified as pro-eating disorder content (and disturbingly, another **13%** displayed self-harm imagery).



02. Content moderation systems can create risks.

We reported explicitly pro-eating disorder content and found that platforms failed to remove this content as they claim to in their policies, meaning it stayed visible on their platform in violation of their guidelines.

Specifically:

- On TikTok, **15.5%** of 110 reported posts were removed;
- On Instagram, **6.3%** of 175 reported posts were removed;
- On X, **6.0%** of 100 reported posts were removed.

03. Ad approval systems can create risks.

We created 12 'fake' ads that promoted dangerous weight loss techniques and behaviours. We tested to see if these ads would be approved to run, and they were. This means dangerous behaviours can be promoted in paid-for advertising. (Requests to run ads were withdrawn after approval or rejection, so no dangerous advertising was published as a result of this experiment.)

Specifically:

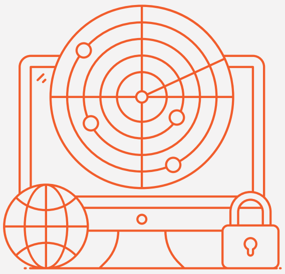
- On TikTok, **100%** of the ads were approved to run;
- On Facebook, **83%** of the ads were approved to run;
- On Google, **75%** of the ads were approved to run.

04. Ad management systems can create risks.

We investigated how platforms allow advertisers to target users, and found that it is possible to target people who may be interested in pro-eating disorder content.

Specifically;

- On TikTok: End-users who interact with pro-eating disorder content on TikTok, download advertisers' eating disorder apps or visit their websites **can be targeted**;
- On Meta: End-users who interact with pro-eating disorder content on Meta, download advertisers' eating disorder apps or visit their websites **can be targeted;**
- On X: End-users who follow pro-eating disorder accounts, or 'look' like them, **can be targeted**;
- On Google: End-users who search specific words or combinations of words (including pro-eating disorder words), watch pro-eating disorder YouTube channels and probably those who download eating disorder and mental health apps **can be targeted**.



Risks to Australians' safety and wellbeing are manifesting in numerous online systems and a regulatory framework needs to incentivise platforms to proactively identify and comprehensively mitigate these risks.

To achieve this, we recommend that:

- The Basic Online Safety Expectations be amended to include additional expectations that online service providers take reasonable steps regarding all systems and elements involved in the operation of their service. Of the four systems explored in this research, only recommender systems would be covered by the current proposals. Many others have not been investigated, such as search systems or engagement features, that will likewise create risks. Safety expectations should be broad and cover all systems and elements deployed by digital platforms.

- The *Online Safety Act* review should implement:
  - An overarching duty of care on platforms;
  - Risk assessments and risk mitigation obligations across all systems and elements;
  - Meaningful transparency measures to make publicly visible the risks and mitigation measures created by systems and elements, and;
  - Strong accountability and enforcement mechanisms.

CONTENTS

Reset.
AUSTRALIA

Reset.Tech Australia is an independent, non-partisan policy research lab committed to driving public policy advocacy, research and civic engagement to strengthen our democracy within the context of technology. We are the Australian affiliate of Reset, a global initiative working to counter digital threats to democracy.

INTRODUCTION

What are systems and elements?

This report explores how platforms’ systems and elements create risks. When we talk about ‘systems and elements’, we are referring to the specific processes and features platforms build into their products, such as recommender systems, search features, ‘like’ buttons, interactive features (like chats and comment functionality) and so on. Systems and elements are not organic, and each system and element deployed on a platform represents a deliberate design choice made by a platform. They will have been carefully crafted by engineers, designers and UX experts and are entirely within the control of the platform.

Legislation is emerging all around the world that aims to improve user safety online by making platforms accountable for the risks their systems and elements create, such as the risks of amplifying harmful content through recommender algorithms or failing to prevent it in paid-for ads. For example, the EU’s *Digital Services Act* asks platforms to “focus on the systems or other elements that may contribute to the risks”[1] and the UK’s *Online Safety Act* places duties of care on platforms to keep users safe “across all areas of a service, including the way it is designed, operated and used as well as content present on the service”.[2]

Australia is currently reviewing our *Online Safety Act*, and the Basic Online Safety Expectations within. There are proposals to introduce safety requirements around a number of systems and elements—for example, generative AI capabilities, recommender systems, user controls, enforcement of terms of use, complaints and reporting systems—but many others are not included in the proposals.

This research aims to highlight why a systemic approach to online systems and the digital risks they create is required. It encourages the government to consider introducing requirements that digital platforms take reasonable steps to ensure users’ safety across all systems and elements of their service, in keeping with emerging global norms, and place a duty of care on platforms for user safety. The need for this is highlighted by looking at how four different systems can be involved in creating online risks regarding pro-eating disorder content and behaviours.

This research explores the role of recommender systems (which are included in the proposals for the Basic Online Safety Expectations) but also the role of content moderation systems, ad approval systems and ad management systems in perpetuating risk. It explores a range of platforms, including TikTok, Instagram, X and Google. Pro-eating disorder content is used here as an example of an online risk, but the argument applies to all other sorts of content of concern such as drug, alcohol, gambling or self-harm. Given the vast array of potential risks arising from online systems, a comprehensive approach that probes digital platforms at a systemic level is arguably needed to ensure all risks to Australians’ safety are adequately mitigated.

The relationship between eating disorders and social media content consumption

Eating disorders and disordered eating are significant public health issues in Australia. Around one million Australians experience an eating disorder each year, roughly 4% of the population, with a third of Australian teenagers showing signs of disordered eating (but not meeting the threshold for being diagnosed with an eating disorder).[3] Eating disorders can have serious and complicated impacts for those who experience them, creating a range of other clinical health issues and sometimes leading to death.[4]

An emerging body of evidence has connected social media use with eating disorders and disordered eating. Although the onset of an eating disorder in any individual will be a complex process involving multiple factors, how social media is used appears to be a risk factor.

For example:

› Australian studies have shown that the use of specific features on social media connects with eating disorders. For example, one study investigating engagement associated with 'friending' found that a higher number of Facebook friends was associated with a 'drive for thinness' in Australian teenagers.[5] Another study found that greater engagement in photo activities on Facebook, but not general Facebook use, was associated with greater internalisation of the 'ideal body' and body monitoring;[6]

› Australian research suggests that the platforms young people use, and the extent to which that are used, also matter. Having more social media accounts was associated with higher disordered eating attitudes and behaviours. Further, time spent on social media was associated with higher disordered eating behaviours in girls, and this varied by platform;[7]

› A global systematic review found that time spent on social media was related to body image concerns and the possible development or perpetuation of eating disorders. It also found that specific functionalities of social media platforms matter. The types of interactions and engagement made available on digital platforms—such as being able to read and write comments, the ability to share photos or post selfies—were connected to feelings of control over and bodily dissatisfaction.[8]

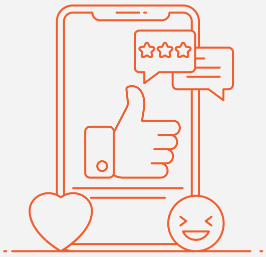
This research explores the connection between eating disorders and digital platform functionalities the 'inverse way'. Rather than starting with eating disorder symptomology and prevalence, and exploring correlations or causal links to platform use or functionalities, we start with digital platforms and explore how their functionalities create risks. In this way, we hope to contribute analysis highlighting the connections between platforms' functionalities—as actively designed into platforms' systems and elements—and risks of eating disorders.

We also hope this research shows that systemic regulation of digital platforms' vast and powerful functionalities is a key way to mitigate risks to the public, including public health risks. This report explores the potential of the *Online Safety Act* as a primary vehicle for risk-based and systemic platform regulation. Appendix 4 identifies the tools available in the *Privacy Act*.

1. CONTENT RECOMMENDER SYSTEMS

Content recommender systems are the elements of a platform that organise, prioritise and ultimately promote content onto users' feeds. They generate users' feeds, and decide what content is recommended in response to various search terms and functionalities. These algorithmic systems are often complex and shrouded in mystery and are extremely powerful. At one stage, YouTube's executives claimed that up to 70% of the content consumed on the platform was driven by their recommender system.[9] Content recommender systems can be key drivers of risk; where they fail, they can promote and recommend harmful content to users—including vulnerable users. They can amplify potential harms and affect a wider set of users.

Platforms' community guidelines prohibit and discourage users from 'posting' pro-eating disorder content (see Appendix 1). Some platforms claim to 'demote' this content; that is, they claim to use their recommender systems to restrict the spread of some content and decrease harm. Other platforms simply claim this content is not allowed, without necessarily explaining their approach to handling this content if posted.



The experiment

Working on TikTok, Instagram and X (formerly Twitter), we set out to see if content moderation systems would promote pro-eating disorder content to users' feeds and amplify the harms, or if they would demote it and reduce risks. To do this, we set up an account notionally belonging to a 16-year-old Australian on each platform. We primed each account by 'liking', 'hearting', or on TikTok, 'rewatching' and 'liking' 50 pieces of pro-eating disorder content on each account.

We then tracked the next 355 pieces of content that our fake child's account was recommended. This was in the 'For You' feed on TikTok, the 'Search' feed on Instagram and the 'For You' feed on X. We counted how much of this recommended content was pro-eating disorder content (see Appendix 2 for details and definitions of what was counted). On X, the 'For You' feed was refreshed every 10 posts. On Instagram, the 'Search' feed was refreshed after every 10 images. TikTok's feed automatically refreshes.

The findings are as follows:

- On TikTok, we found no evidence that TikTok's content recommender system will promote pro-eating disorder content, suggesting that they have placed safeguards on their algorithm.
  - 0% of the content could be classified as pro-eating disorder content, despite 'priming' the account by rewatching and liking pro-eating disorder content (see Figure 1).
  - TikTok's processes may have gone further and actively 'unliked' the pro-eating disorder content we 'liked' on our account. Unliking videos is a commonly reported glitch on TikTok,[10] but there are some reports that TikTok 'unlikes' videos with inappropriate content.[11] For this experiment, the account should still be considered primed because we double-watched 50 pro-eating disorder videos and evidence suggests that the length of video views is the most powerful data used by TikTok's content recommender system.[12]
  - We rested this account for one week but still were not recommended any content that could be classified as pro-eating disorder content.
  - We note that this experiment involved only one account, so TikTok's recommender system may still recommend pro-eating disorder content to other accounts not using this controlled methodology. However, the evidence here seems to suggest that steps have been taken to limit this capacity.
- On Instagram, we found evidence that Instagram's recommender systems will promote pro-eating disorder content, but that their algorithm may have some safeguards in place that function to prevent its immediate promotion and prevent it from comprising the majority of a search feed.
  - 23% of the content could be classified as pro-eating disorder content (see Figures 2 and 3).
  - No pro-eating disorder content was initially recommended to our account. We opened the account, noted that no pro-eating disorder content was recommended in the first 50 posts, and then rested the account for one week (i.e. we did not log in for 7 days). After one week, we opened the account again and found some pro-eating disorder content appearing on the Search feed. This change was immediate; the first two posts after the 7 day reset were pro-eating disorder content.
  - We note that this experiment involved only one account, so Instagram's recommender system may still recommend pro-eating disorder content to other accounts immediately or at greater quantities. However, the evidence here seems to suggest that steps have been taken to limit this capacity.

› On X, we found evidence that X's recommender systems will promote pro-eating disorder content, and we found no evidence of safeguards on their algorithms.

- 67% of the content could be classified as pro-eating disorder content (see Figures 4 and 5).
- This content was received immediately; the first four recommended posts contained pro-eating disorder content. However, it is important to note that this account was primed, so this might not be the experience for an unprimed child's account.
- Worryingly, the recommender systems also promoted other types of harmful content. After 37 pieces of content, the recommender systems began additionally recommending explicit self-harm content (such as images of self-harm cuts; see Figure 6). By the end of the experiment, 13% of content recommended by X's algorithm depicted self-harm.
- In total, 80% of the content served by the algorithm could be described as either pro-eating disorder content or explicitly depicted self-harm. Towards the end of the experiment, the recommender systems were recommending almost entirely pro-eating disorder or self-harm posts; of the final 10 posts, all were classified as pro-eating disorder or self-harm content.

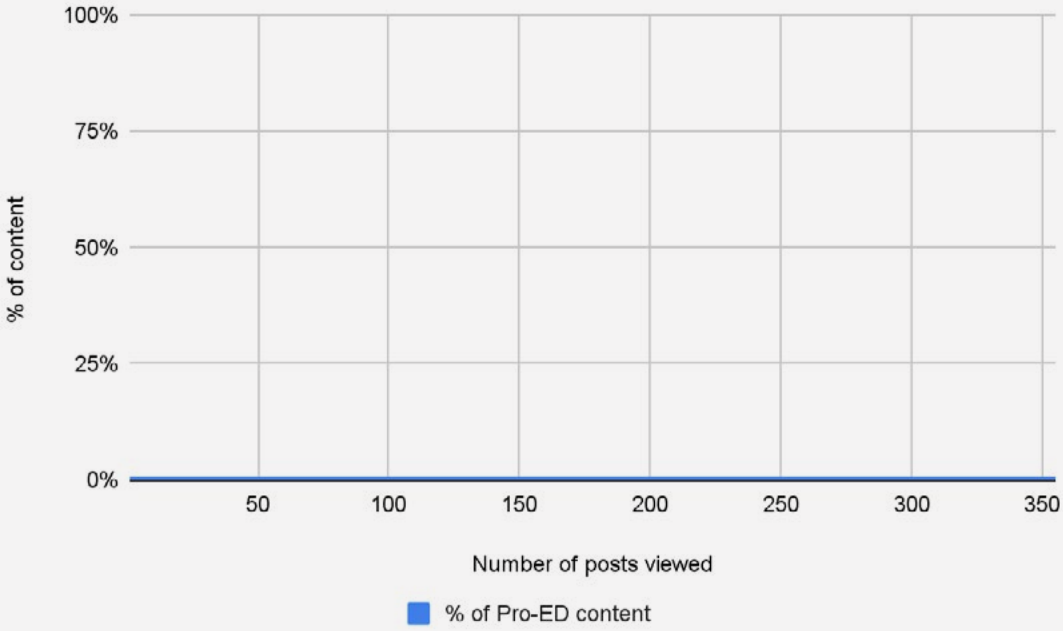


Figure 1: The percentage of pro-eating disorder content recommended to our fake child's account on TikTok (n=355 posts)

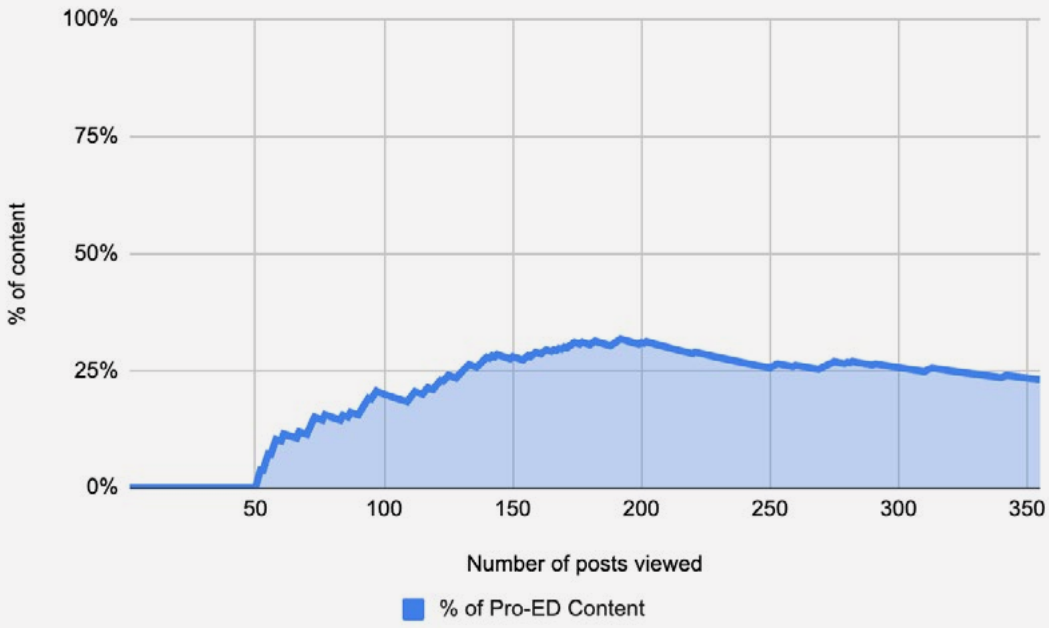
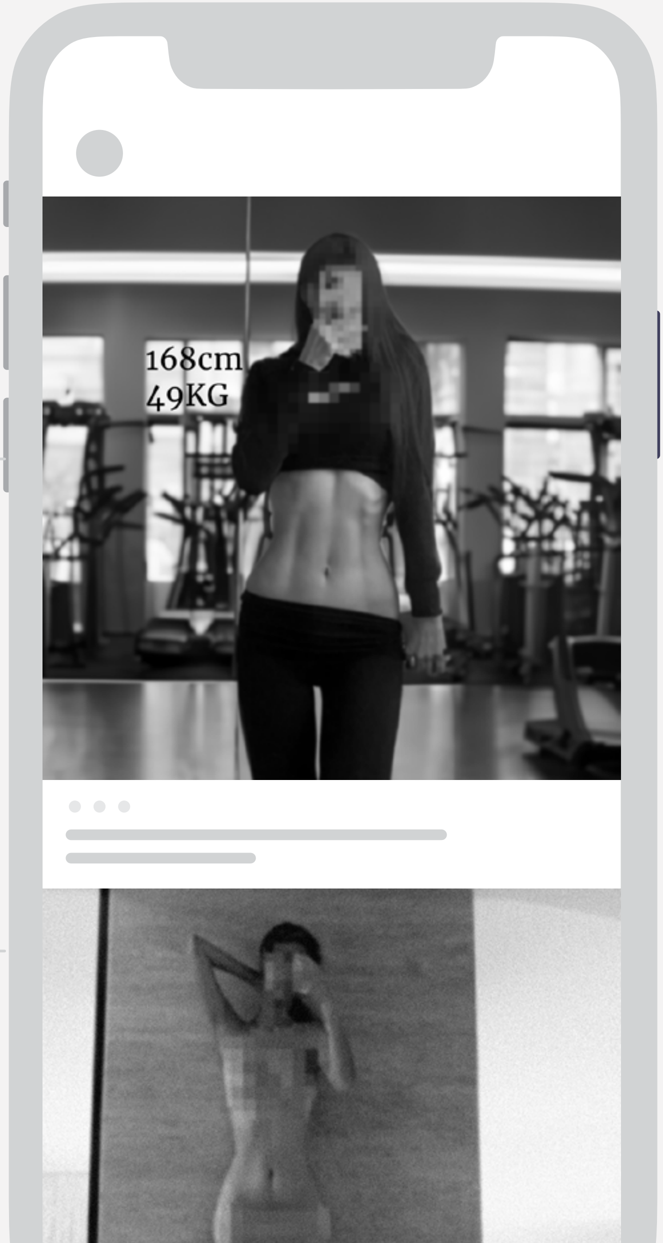


Figure 2: The percentage of pro-eating disorder content recommended to our fake child's account on Instagram (n=355 posts)

Figure 3: Examples of pro-eating disorder content recommended to our fake child's account on Instagram



The image is a relfection of the person or female taking a photo in front of a mirror. The smart phone covers the face and the midrift is exposed. Height and weight is overlayed as text.

An image of a person or female clothed in under garments taking a photo in front of a mirror.

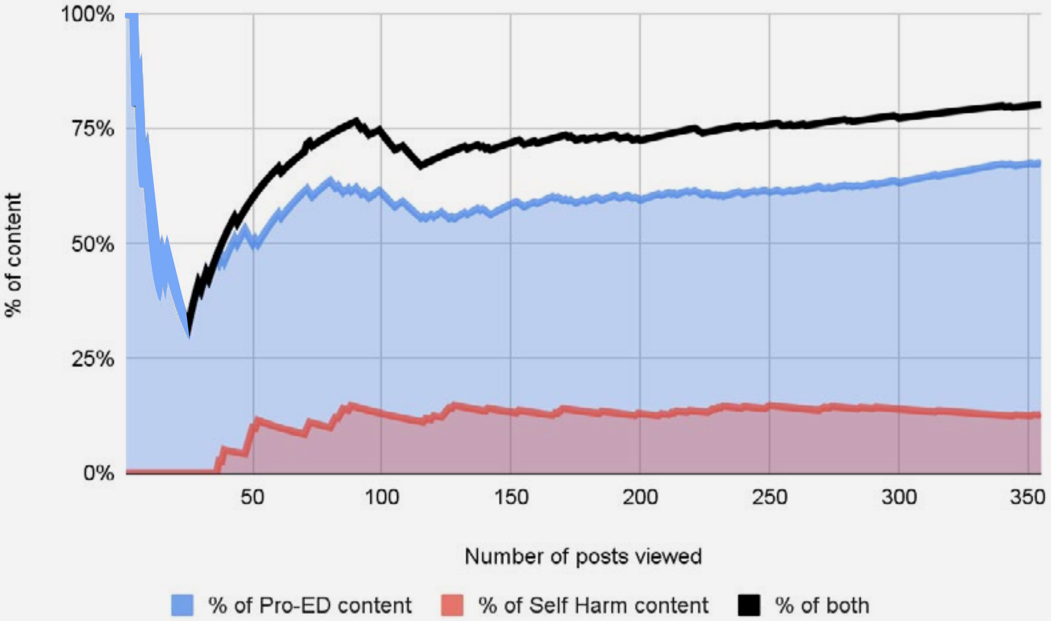
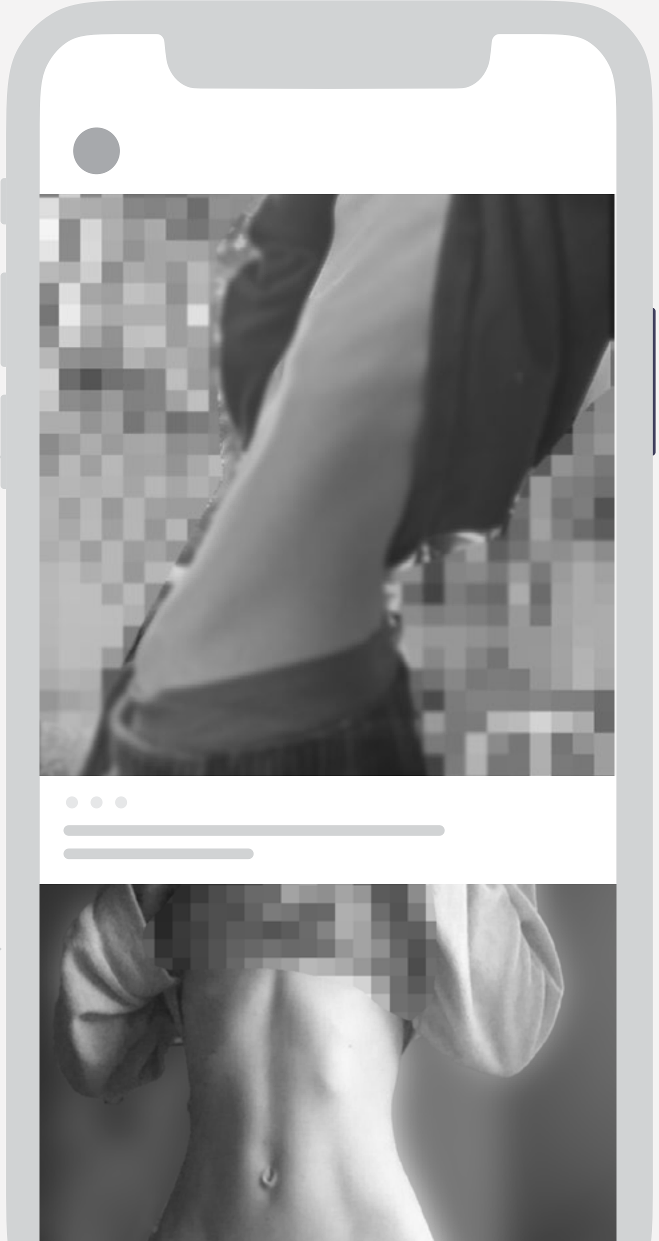


Figure 4: The percentage of pro-eating disorder content recommended to our fake child's account on X (n= 355 posts)

Figure 5: Examples of pro-eating disorder content recommended to our fake child's account on X



The image is of an unidentifiable persons midrift.

The image is of a person or female clothed in under garments with their midrift exposed.

These findings highlight a systemic issue

Content recommender systems have been routinely found to promote content that risks mental or physical injury, such as eating disorder content,[13] age-inappropriate violent, extremist content,[14] or misogynistic content.[15] Friend or follower recommender systems can also be part of the problem. Research has shown how friend recommender systems create 'eating disorder bubbles',[16] and they can also promote connections between children's and adult's accounts that create grooming risks.[17] There is also evidence that it is possible to train algorithms not to do this,[18] so these affordances should be considered a design choice.

What this shows

Content moderation systems routinely promote risky content, which amplifies the potential harms for vulnerable users online, but they can also be designed in ways that minimise risks. Placing obligations on platforms to ensure their content recommender systems are safe and work to *demote* harmful content—rather than promote it—would help create a comprehensive regulatory framework that protects users.

The proposed reforms to the Basic Online Safety Expectations include proposals to ensure basic safety obligations regarding recommender systems are in place, and this research would support this inclusion.

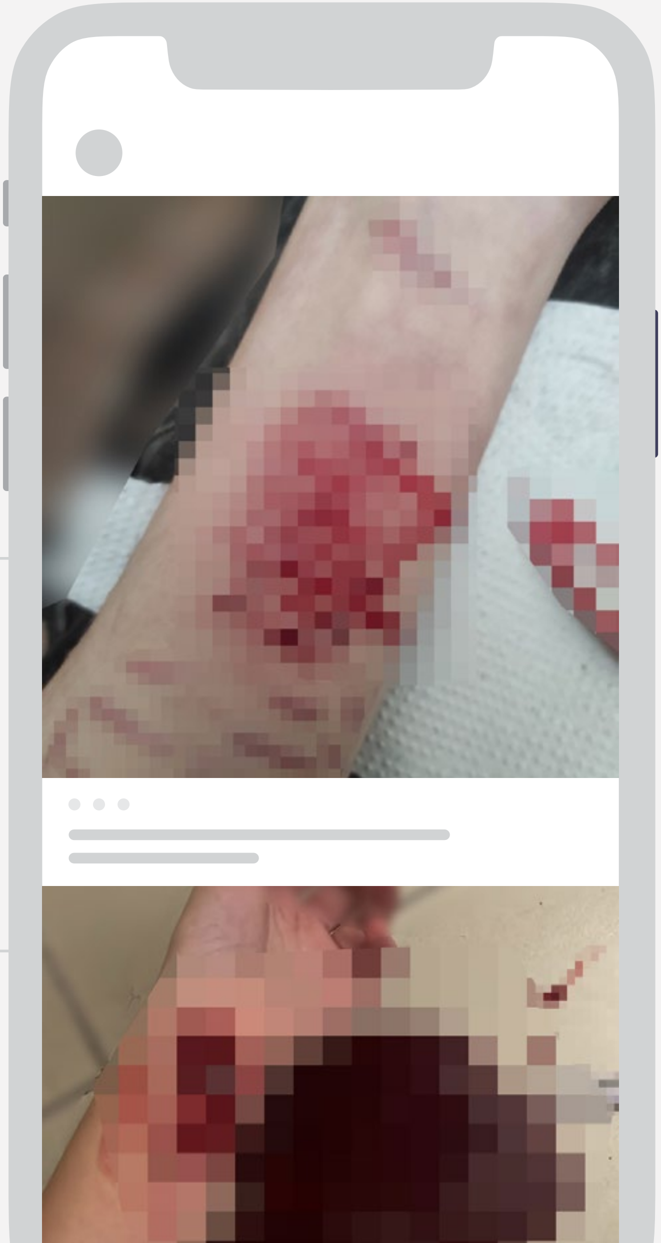


Figure 6: Examples of self-harm content recommended to our fake child's account on X

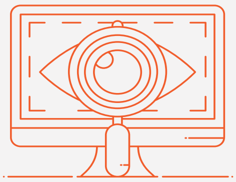
The image is of a persons wrist who has self harmed.

The image is of a persons wrist who has self harmed in a pool of blood.

2. CONTENT MODERATION SYSTEMS

Content moderation systems are an integral part of ensuring safety on any service that hosts user-generated content. They are the systems that ensure that online service providers are able to detect, classify and respond to content on their platform that breaches guidelines. This includes systems that proactively detect violative content and systems that respond to user-reports of violative content. Once a platform is aware of a piece of violative content, there are many actions it can take. They could remove the content, apply a ‘warning label’ or ‘sensitivity filter’ to it, reduce its visibility or demote it within their recommender systems. Content moderation systems can be key drivers of risk; where they fail, harmful content remains visible on their platform.

When it comes to pro-eating disorder content on social media, most platforms have policies that outline that they remove harmful content once they become aware of it (see Appendix 1). Failure to implement this adequately would result in risky pro-eating disorder content remaining accessible despite user-reporting.



The experiment

Working on TikTok, Instagram and X, we tested to see if harmful content was in fact removed from the platforms once the platforms became aware of it. To do this, we identified pro-eating disorder content using a codebook (see Appendix 2) and monitored it for one week to see if it was detected or removed by the platform. We then reported it to the platform using the user-reporting function and monitored it for another week to see if it was subsequently removed.

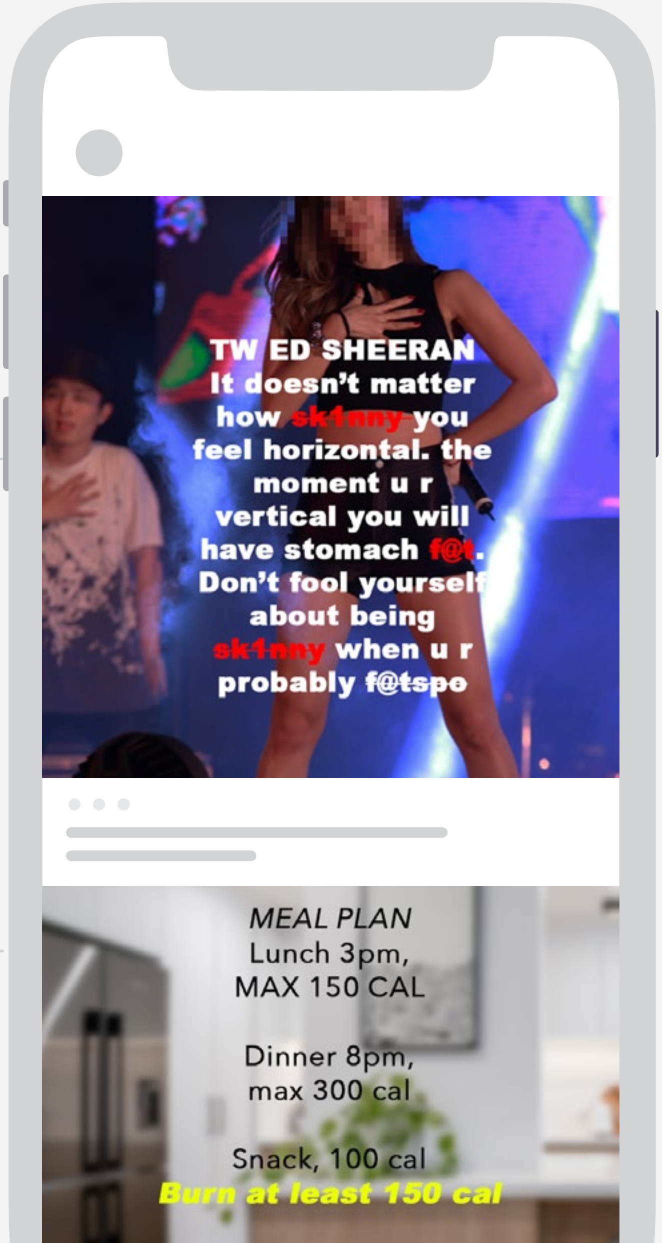
The findings are as follows:

- On TikTok, we identified 110 pieces of pro-eating disorder content to monitor and report. Content was coded against the codebook (see Appendix 2) to confirm it was harmful by two researchers. We found that reporting content had minimal effect on take down rates, and no impact on labelling rates (see Figure 7).
  - Before reporting, three pieces of content (2.7%) had become unavailable. All three pieces of content were still available to users logged in as an adult, but they had become unavailable publicly and for under 18-year-olds, so we have counted these as becoming unavailable. No content featured a warning label.
  - After reporting, an additional 14 pieces of content (12.7%) became unavailable, and no pieces of content featured warning labels.
  - In total, 17 pieces of content (15.5%) became unavailable overall. Three of these were still available to users logged in as an adult, but they had become unavailable publicly and for under 18-year-olds, so we have counted these as becoming unavailable. Fourteen of them were entirely unavailable. No content featured a warning label.
- On Instagram, we identified 175 pieces of pro-eating disorder content to monitor and report. We had an eating disorder academic verify that each piece of content included was harmful and correlated to the codebook. We found that reporting content had minimal effect on take-down rates and no impact on labelling rates (see Figure 7).
  - Before reporting, two pieces of content (1.1%) had become unavailable. Both pieces had become unavailable because the accounts had been switched to private (meaning we cannot ascertain whether the content became unavailable because it was taken down by Instagram or because the user had changed their account settings). We have counted these as becoming unavailable. No content featured a warning label.
  - After reporting, an additional nine pieces of content (5.1%) had become unavailable, and no pieces of content featured warning labels.
  - In total, 11 pieces of content (6.3%) became unavailable overall. Two of these were unavailable because the account had been switched to private (meaning we cannot ascertain whether the content became unavailable because it was taken down by Instagram or because the user had changed their account settings). Nine were completely unavailable. No content featured a warning label.
- On X, we identified 100 pieces of pro-eating disorder content to monitor and report. Content was coded against a codebook (see Appendix 2) to confirm it was harmful by two researchers. We found that reporting content had minimal effect on take-down rates and no impact on labelling rates (see Figure 7).
  - Before reporting, zero pieces of content (0%) had become unavailable, and no pieces of content featured warning labels.
  - After reporting, six pieces of content (6.0%) became unavailable. No content featured a warning label.
  - In total, six pieces (6.0%) became unavailable. X provides some information about why content becomes unavailable. Two pieces became unavailable because they ‘violated the X rules’, one piece because an account was suspended, one piece because the account changed settings to limit who could view the content, and two pieces lead to ‘something went wrong’ error pages. No content featured warning labels or blurring filters on images.

Over two weeks of monitoring	TikTok (n=110)	Instagram (n= 175)	X (n=100)
Efficacy of platforms' proactive detection and response rates (i.e. pre-reporting removal rate). This is the % of content that was removed during the week before we reported it. Content may have been reported by other users, and it is often not clear why content was removed (e.g. users may have deleted the content or their accounts, moved to private, or platforms may have deleted it). However, this represents the best estimate of organic removal rate.	2.7%	1.1%	0%
Efficacy of user-reporting systems (i.e. post-reporting removal rate). This is the % of content that was removed one week after reporting and is the best estimate of the impact of user reporting.	12.7%	5.1%	6.0%
Efficacy of content moderation systems overall (i.e. total content removal rate). This is the total amount of content that was removed overall and is the best estimate of the impact of content moderation systems overall.	**15.5%**	**6.3%**	**6.0%**

Figure 7: The effect of reporting pro-eating disorder content on removal rates, by platform (Australian experiment Feb 2024)

Figure 8: Examples of pro-eating disorder content that was not removed from TikTok despite reporting



TW ED SHEERAN
It doesn't matter how sk1nny you feel horizontal. the moment u r vertical you will have stomach f@t. Don't fool yourself about being sk1nny when u r probably f@tspo

MEAL PLAN
Lunch 3pm,
MAX 150 CAL

Dinner 8pm,
max 300 cal

Snack, 100 cal
Burn at least 150 cal

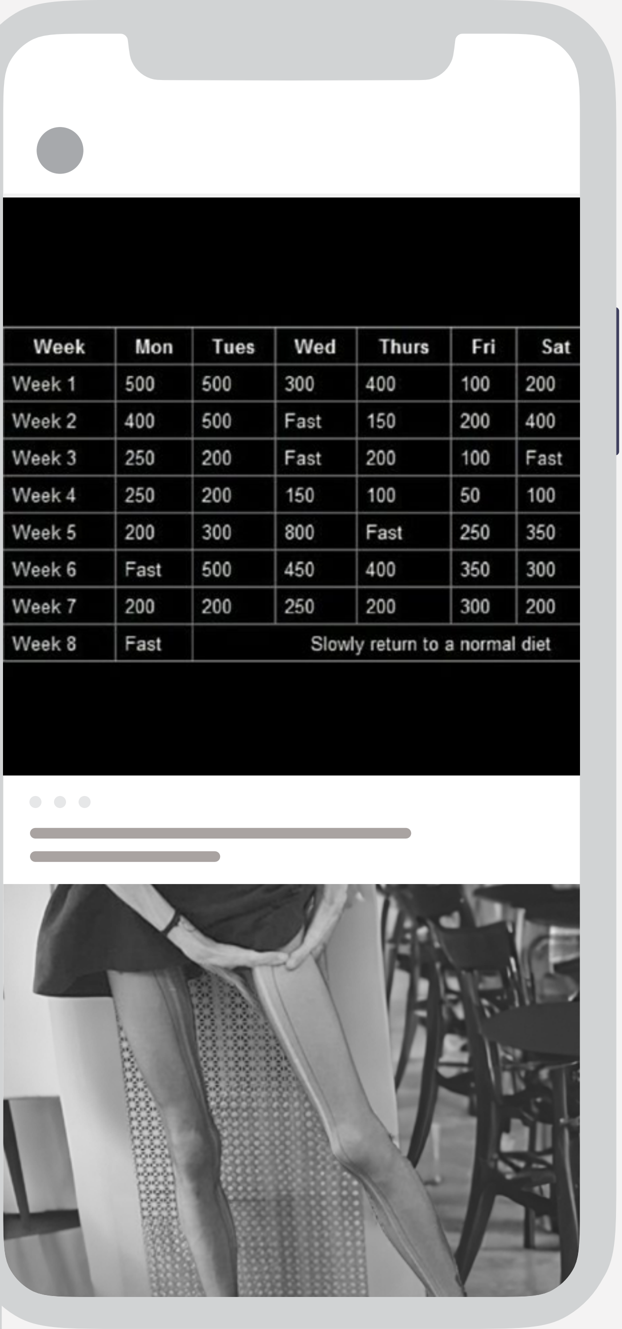
Still-frame taken from video is of choregraphed dancing with text overlayed.

Text overlayed background image of a kitchen

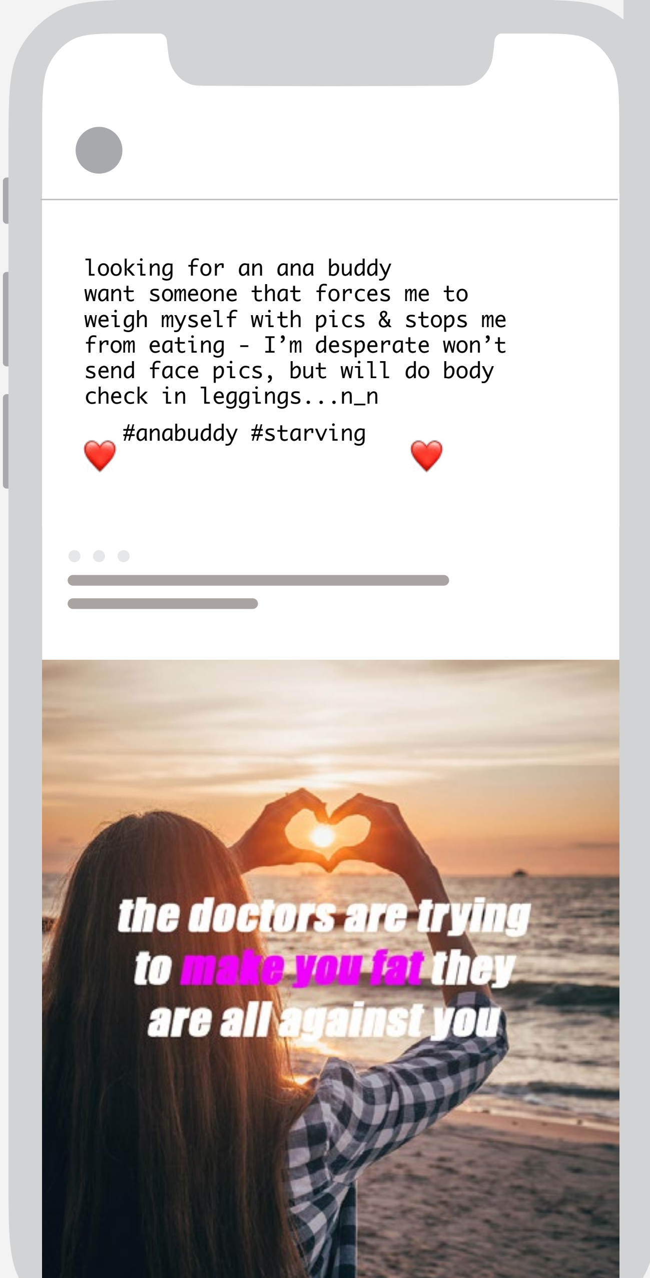
Figure 9: Examples of pro-eating disorder content that was not removed from Instagram despite reporting

Image is a static graphic chart representation of calorie counting per day and week.

The image is of an unidentifiable persons legs with both persons hands, pinky to thumb, encircling their thigh.



Week	Mon	Tues	Wed	Thurs	Fri	Sat
Week 1	500	500	300	400	100	200
Week 2	400	500	Fast	150	200	400
Week 3	250	200	Fast	200	100	Fast
Week 4	250	200	150	100	50	100
Week 5	200	300	800	Fast	250	350
Week 6	Fast	500	450	400	350	300
Week 7	200	200	250	200	300	200
Week 8	Fast	Slowly return to a normal diet				



looking for an ana buddy
want someone that forces me to
weigh myself with pics & stops me
from eating - I'm desperate won't
send face pics, but will do body
check in leggings...n_n

❤️ #anabuddy #starving ❤️

Figure 10: Examples of pro-eating disorder content that was not removed from TikTok despite reporting

Text post on X.

The image is of person making a heart shape around a sunset with their hands with text overlayed that reads: “the doctors are trying to make you fat. They are all against you”

These findings highlight a systemic issue

We undertook similar research across Europe in November 2023 and found similar results.[19] At best, 17% of pro-eating disorder content was removed on TikTok, with Instagram removing 10% and X removing 9% (see Figure 11).

	TikTok (n= 107)	Instagram (n=125)	X (n=111)
Efficacy of platforms' proactive detection and response rates (i.e. pre-reporting removal rate). This is the % of content that was removed before reporting. Content may have been reported by other users, and it is often not clear why content was removed (e.g. users may have deleted the content or their accounts, moved to private, or platforms may have deleted it). However, this represents the best estimate of organic removal rate.	5.6%	0%	2.7%
Efficacy of user-reporting systems (i.e. post-reporting removal rate). This is the % of content that was removed after reporting and is the best estimate of the impact of user reporting.	11.2%	10.4%	6.3%
Efficacy of content moderation systems overall (i.e. total content removal rate). This is the total amount of content that was removed overall and is the best estimate of the impact of content moderation systems overall.	**16.8%**	**10.4%**	**9.0%**

Figure 11: The effect of reporting pro-eating disorder content on removal rates, by platform (EU experiment Nov 2023)

In the EU, platforms have introduced new user-reporting flows and procedures to meet more stringent requirements under their *Digital Services Act*. This does not apply to Australian users.

What this shows

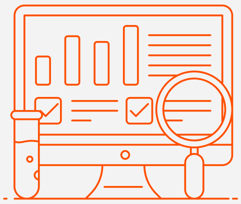
Platforms may have strong policies against hosting pro-eating disorder content, but these policies are not reliably enforced. Content moderation systems routinely fail, which leaves harmful content available online. Placing obligations on platforms to ensure their content moderation systems are effective and work to enforce their community guidelines would be helpful in creating a comprehensive regulatory framework that protects users.

The proposed reforms to the Basic Online Safety Expectations do not include proposals to ensure basic safety obligations regarding content moderation systems are in place.
We strongly encourage the inclusion of content moderation systems in the Expectations.

3. AD APPROVAL SYSTEMS

Ad approval systems determine which ads can run on a platform and are meant to detect and block violative ads, in accordance with each platform's advertising policies. These systems are often automated, with technology used to detect ads that potentially breach their guidelines. Others claim to be human-moderated, and some use a combination of the two ('human in the loop').[20] Ad approval systems can be key drivers of risk; where they fail, they allow harmful content to reach a wide audience via paid-for advertising.

Most platforms have policies against placing harmful content in paid-for advertising, which includes dangerous weight loss ads (see Appendix 3). Failure to implement this adequately would result in risky ads being approved, meaning that a platform's paid-for advertising system is vulnerable to be hijacked and cause harm.



The experiment

Working on TikTok, Facebook and Google, we set out to see if platforms' ad approval systems blocked pro-eating disorder content in paid-for advertising. To do this, we set up ad-enabled accounts on each platform, as per the requirements of each platform. We then developed 12 fake ads that included dangerous weight loss techniques, working with an eating disorder academic to ensure that all of the ads could be considered harmful. We then put these fake ads through each platform's ad approval systems. We monitored these ads to see if they were approved or not, and once they had been assessed by a platform, we cancelled each ad. To be clear, no harmful ad was run as a result of this experiment and no one saw these ads outside of the context of this research.

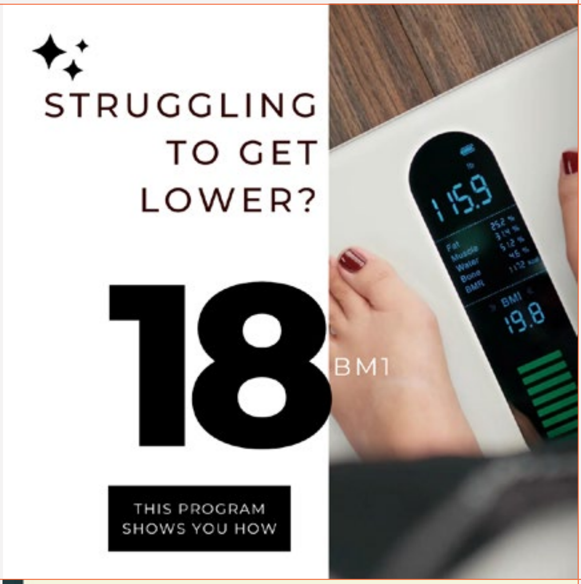
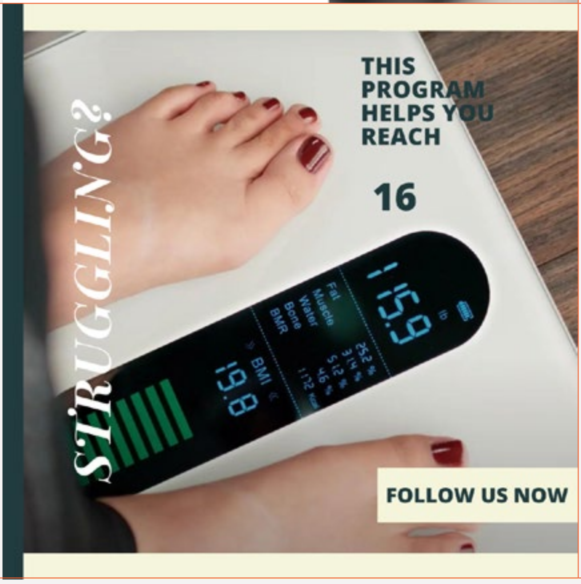
The findings are as follows:

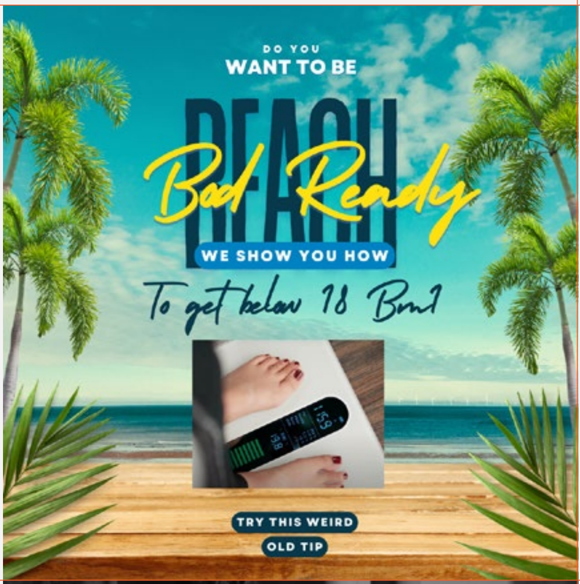
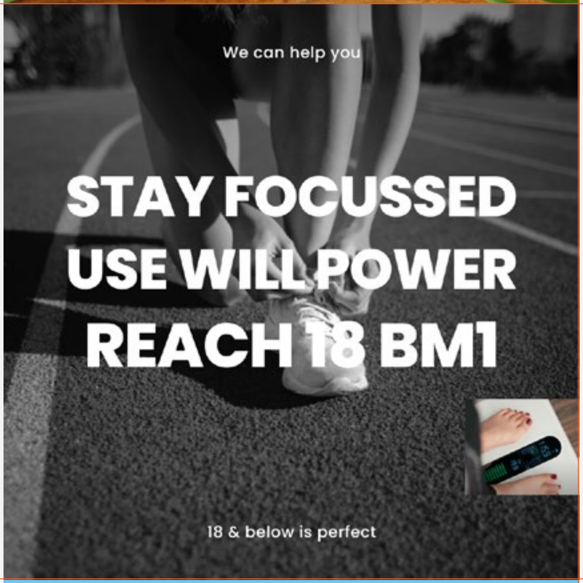
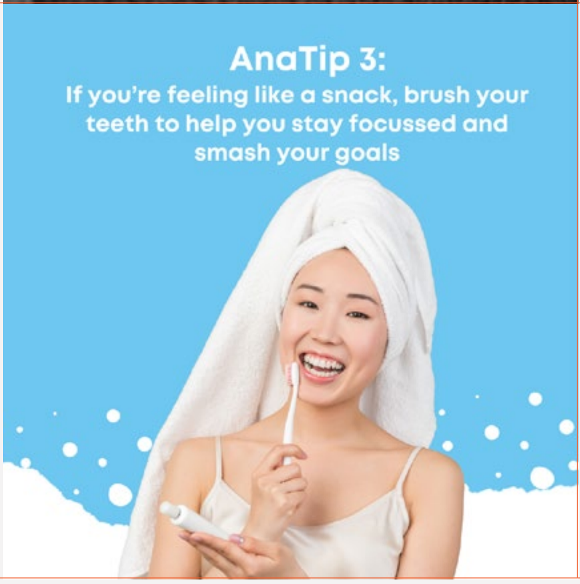
- On TikTok, 100% of ads were approved to run (12 out of 12).
- On Facebook, 83% of ads were approved to run (10 out of 12). Two ads were rejected for violating personal health and appearance rules, but others which also violated the rules were approved.
- On Google, 75% of ads were approved to run (9 out of 12). Three ads were rejected for containing 'clickbait' rather than any pro-eating disorder safety scanning. That is, the reason Google provided for the rejection was the inclusion of text that said, 'This weird old tip...', rather than the substance of the ad. We confirmed this by replacing this text and resubmitting the three ads for approval; the three ads were subsequently approved. In fairness to Google, we note that while only three of the 12 ads were rejected, Google's system worked to prevent clickbait headlines but not dangerous weight loss content.

We were unable to open an ad-enable account on X in the timeline of this experiment.

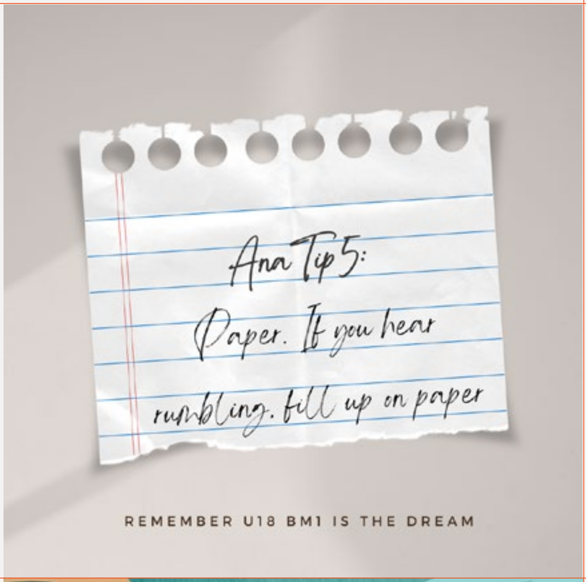
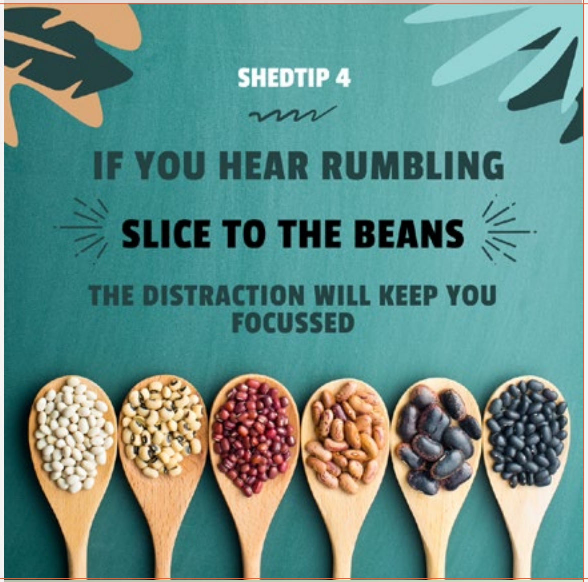
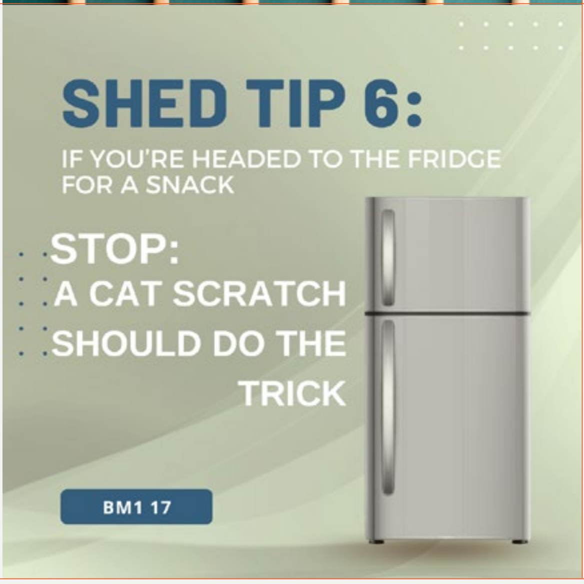
We are aware that these ads did not run to completion, and platforms may have additional safety check that these ads may have been subjected to. However, this would not be not consistent with platforms' own public descriptions about how their advertising approval processes work.[21] It is also inconsistent with previous research undertaken by Reset.Tech Australia, where no 'secondary' approval process was involved in detecting misinformation in advertising.[22]

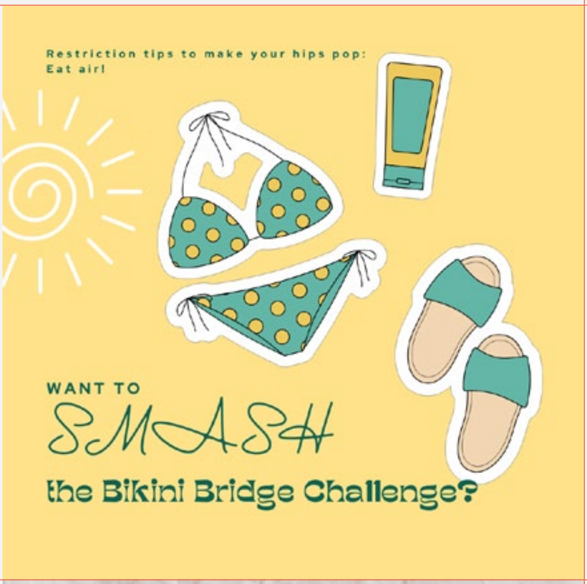
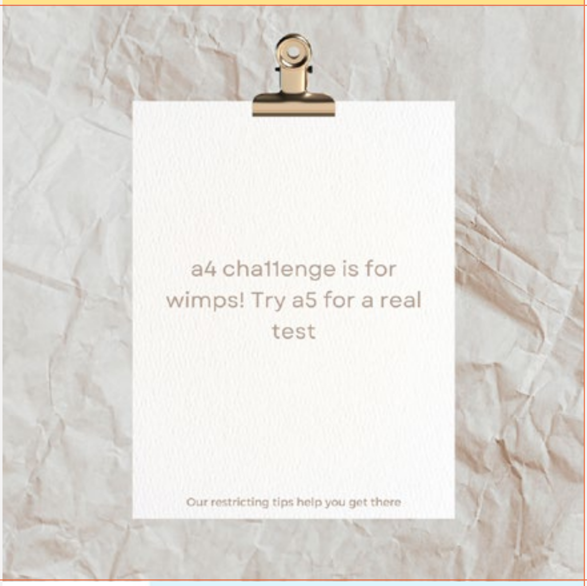
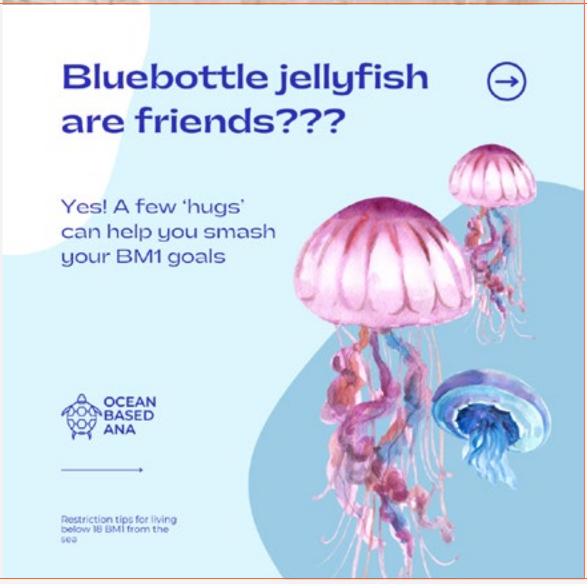
Figure 12: The 12 fake ads that we put forward to test platforms' ad approval systems. Please note ad formats may have varied by platform, but the core elements of the ads (e.g. text and images) were the same

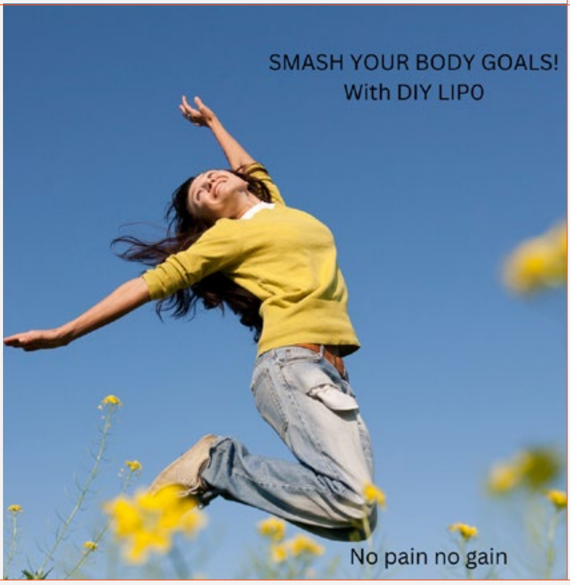
AD NUMBER #		TikTok	Google	Facebook
AD NUMBER 1	 STRUGGLING TO GET LOWER? 18 BM1 THIS PROGRAM SHOWS YOU HOW	✓ ACCEPTED	✗ REJECTED	✗ REJECTED
AD NUMBER 2	 STRUGGLING? THIS PROGRAM HELPS YOU REACH 16 FOLLOW US NOW	✓ ACCEPTED	✓ ACCEPTED	✗ REJECTED

AD NUMBER #		TikTok	Google	Facebook
AD NUMBER 3		✓ ACCEPTED	✗ REJECTED	✓ ACCEPTED
AD NUMBER 4		✓ ACCEPTED	✓ ACCEPTED	✓ ACCEPTED
AD NUMBER 5		✓ ACCEPTED	✓ ACCEPTED	✓ ACCEPTED

AnaTip 3:
If you're feeling like a snack, brush your teeth to help you stay focussed and smash your goals

AD NUMBER #		TikTok	Google	Facebook
AD NUMBER 6	 Ana Tip 5: Paper. If you hear rumbling. fill up on paper REMEMBER U18 BM1 IS THE DREAM	✓ ACCEPTED	✓ ACCEPTED	✓ ACCEPTED
AD NUMBER 7	 SHEDTIP 4 IF YOU HEAR RUMBLING SLICE TO THE BEANS THE DISTRACTION WILL KEEP YOU FOCUSSED	✓ ACCEPTED	✓ ACCEPTED	✓ ACCEPTED
AD NUMBER 8	 SHED TIP 6: IF YOU'RE HEADED TO THE FRIDGE FOR A SNACK STOP: A CAT SCRATCH SHOULD DO THE TRICK BM1 17	✓ ACCEPTED	✓ ACCEPTED	✓ ACCEPTED

AD NUMBER #		TikTok	Google	Facebook
AD NUMBER 9	 Restriction tips to make your hips pop: Eat air! WANT TO SMASH the Bikini Bridge Challenge?	✓ ACCEPTED	✗ REJECTED	✓ ACCEPTED
AD NUMBER 10	 a4 cha11enge is for wimps! Try a5 for a real test Our restricting tips help you get there	✓ ACCEPTED	✓ ACCEPTED	✓ ACCEPTED
AD NUMBER 11	 Bluebottle jellyfish are friends??? Yes! A few 'hugs' can help you smash your BM1 goals OCEAN BASED ANA Restriction tips for living below 18 BMI from the sea	✓ ACCEPTED	✓ ACCEPTED	✓ ACCEPTED

AD NUMBER #		TikTok	Google	Facebook
AD NUMBER 12	 	✓ ACCEPTED	✓ ACCEPTED	✓ ACCEPTED

These findings highlight a systemic issue

These findings are consistent with previous research. For example, in 2021 we experimented to see if we could get ads purporting to contain 'spicy cocktail recipes using only what you can find in your 'rents (parents') liquor cabinet' or ads to help girls 'find your gentleman now 💰' or to win prizes by gambling approved on Instagram to a target 13-17 year old end-users, and we found that these ads were quickly approved.[23] This experiment was repeated internationally, with colleagues finding ads for 'skittles parties' (drug parties) and 'ana tips' (pro-anorexia tips) were all likewise approved.[24] Further suggesting systemic failings, in 2023 we tested the ad approval systems on Facebook, TikTok and X to see if we could get approval to run ads containing electoral process misinformation about the Voice referendum, such as ads suggesting that the referendum was being held on Nov 31st (a non-existent date), or that the referendum had been cancelled or was voluntary or postal. The vast majority (between 70% and 100%) of these ads were approved depending on the platform.[25]

What this shows

Platforms may have strong policies against harmful content in paid-for ads, but these policies are not enforced. Ad approval systems routinely fail, permitting harmful content to be promoted in paid-for advertising. Placing obligations on platforms to ensure their ad approval systems are effective and work to enforce their community guidelines would be helpful in creating a comprehensive regulatory framework that protects users.

The proposed reforms to the Basic Online Safety Expectations do not include proposals to ensure basic safety obligations regarding ad approval systems are in place. We strongly encourage the inclusion of ad approvals systems in the Expectations.

4. AD MANAGEMENT SYSTEMS

Ad management systems manage the process of targeting users with ads. This system starts with personal data collection and analysis and then matches an advertiser with the most 'interested' end-user based on this data. Ad management systems can be key drivers of risk; they create vulnerabilities by using end-users' data and allowing people to be targeted with advertising that they are identified or inferred to be particularly susceptible to. For example, people who exhibit regular gambling behaviours can be singled out to receive persistent and persuasive gambling ads.

The experiment

Working on TikTok, Meta, X and Google, we set out to see if platforms would allow the creation of 'risky targeting'. Specifically, we set out to see if it was possible for advertisers to target end-users interested in pro-eating disorder content. To do this, we used the ad management systems available on each platform to explore the specific ways advertisers could target specific end-users who might be interested in pro-eating disorder content.

We found that platforms would routinely allow advertisers to create risky profiles to target. The findings are as follows:

› On TikTok, end-users who interact with pro-eating disorder content, download advertisers' eating disorder apps or visit their websites can be targeted. Specifically, end-users can be targeted using:
  - Engagements on TikTok, enhanced by including 'lookalike' audiences.
  - Using 'third party' (off TikTok's platform) data such as customer files, website traffic or app downloads and enhanced by including 'lookalike' audiences.
  - To some extent, hashtag targeting.

These processes would not be simple or straightforward, but could be very powerful in accuracy.

› On Meta, end-users who interact with pro-eating disorder content on Meta, download advertisers' eating disorder apps or visit their websites can be targeted. Specifically, end-users can be targeted using:
  - Data derived from Meta sources such as page or account visitor data
  - Using 'third party data' (off Meta's platform) data such as customer lists, website data and app activity
  - To a limited extent, ad interest data if used with a combination of keywords

These processes would not be simple or straightforward, but could be very powerful in accuracy.

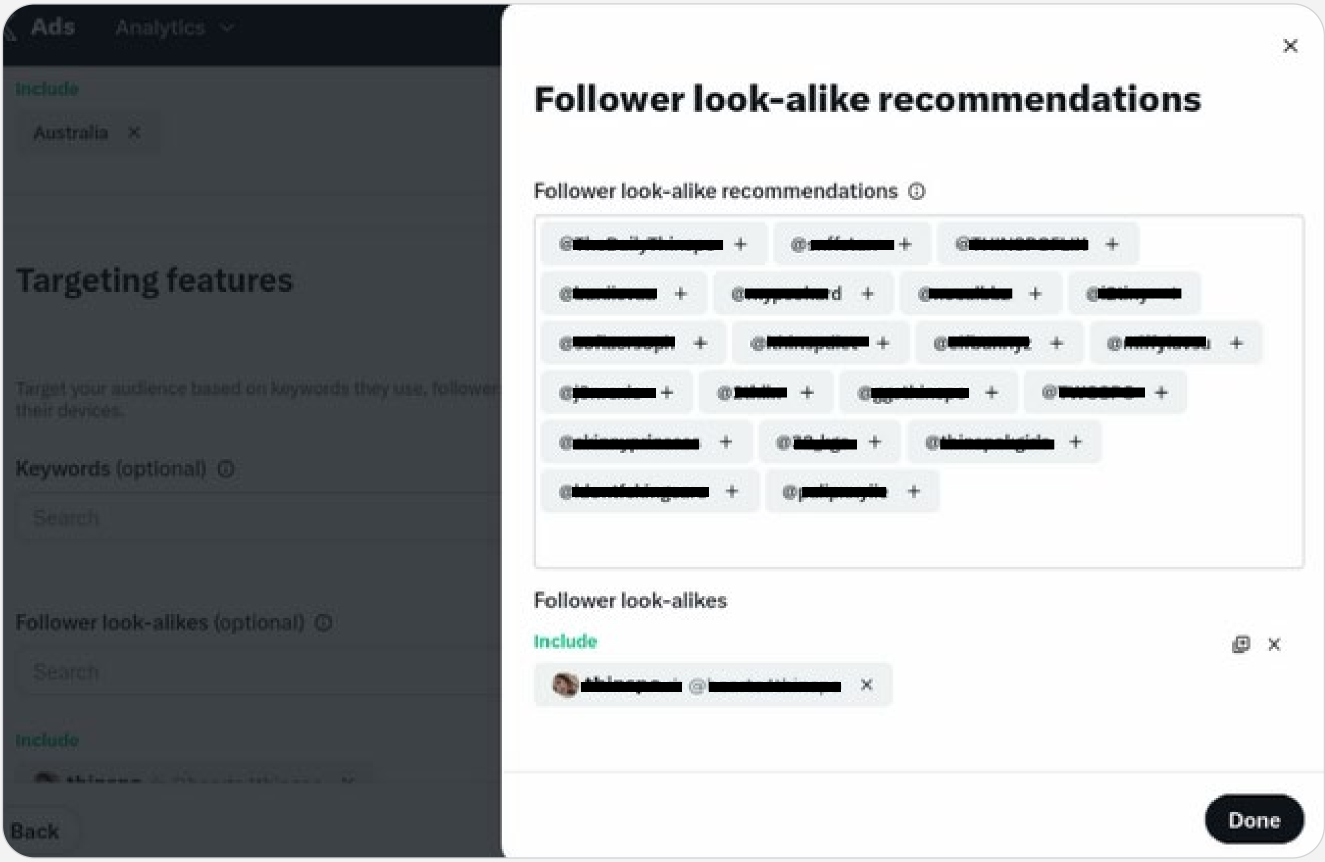
› On X, end-users who interact with pro-eating disorder accounts, or similar accounts, can be targeted. Specifically, end-users can be targeted using:
  - Follower 'lookalikes'; X's ad management system will even recommend pro-eating disorder lookalike audiences to advertisers, once they have entered one pro-eating disorder account to target (see Figure 13)
  - To a lesser extent, key words

These could be comparatively simple and straightforward processes.

› On Google, end-users who search specific keywords or combinations of keywords, and probably those who download general mental health apps can be targeted. Further, ads can be placed in pro-eating disorder YouTube channels. Specifically, end-users can be targeted using:
  - Data about Google keyword searches made;
  - App download data (see Figure 14)
  - YouTube channel data (see Figure 15)

Some of these processes could be comparatively simple or straightforward, and could be very powerful in accuracy.

See Appendix 4 for more details.



×

Follower look-alike recommendations

Follower look-alike recommendations

@[redacted] +
@[redacted] +
@[redacted] +
@[redacted] +
@[redacted] +
@[redacted] +
@[redacted]
@[redacted] +
@[redacted] +
@[redacted] +
@[redacted] +
@[redacted] +
@[redacted] +
@[redacted] +
@[redacted] +
@[redacted] +
@[redacted] +
@[redacted] +
@[redacted] +
@[redacted] +

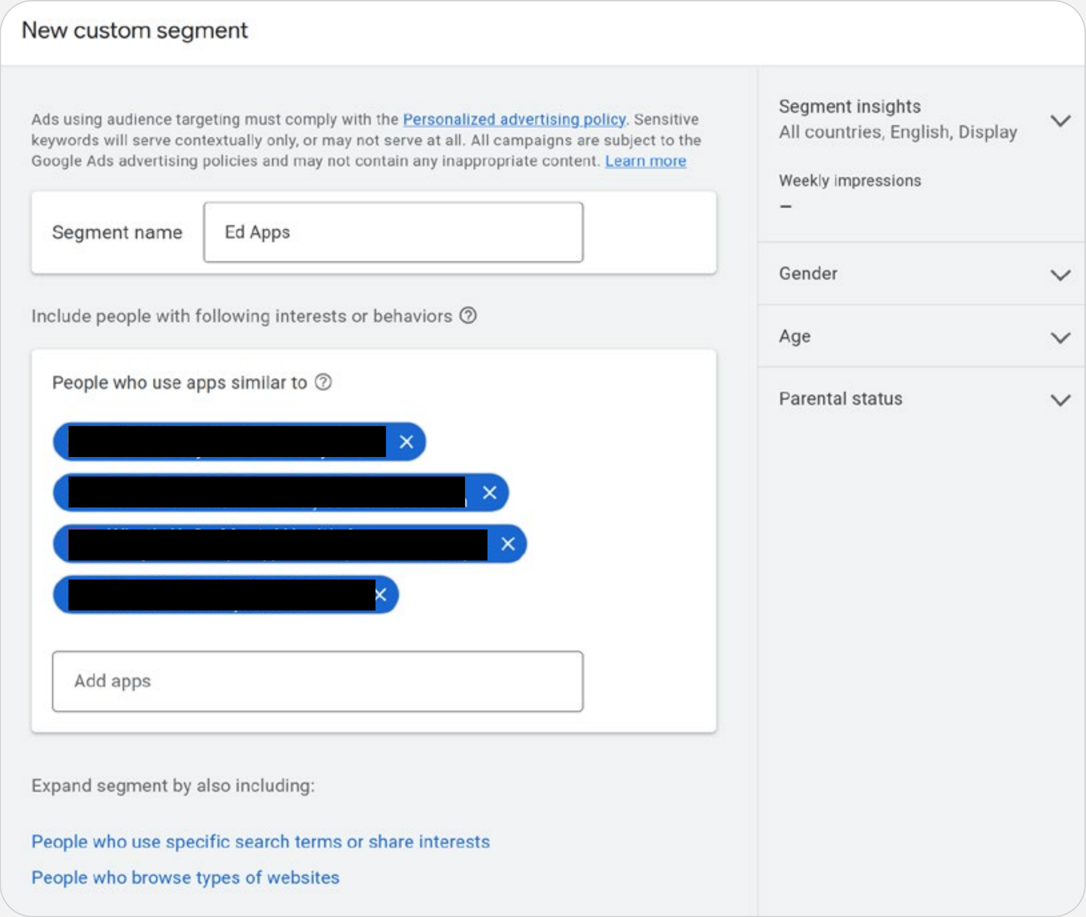
Follower look-alikes

Include

[redacted] @[redacted] ×

Done

Figure 13: X's follower lookalike functionality, showing how it will recommend other pro-eating disorder lookalike audiences to expand the pool of users to target



New custom segment

Ads using audience targeting must comply with the Personalized advertising policy. Sensitive keywords will serve contextually only, or may not serve at all. All campaigns are subject to the Google Ads advertising policies and may not contain any inappropriate content. Learn more

Segment name: Ed Apps

Include people with following interests or behaviors

People who use apps similar to

[redacted] ×

[redacted] ×

[redacted] ×

[redacted] ×

Add apps

Expand segment by also including:

People who use specific search terms or share interests

People who browse types of websites

Segment insights
All countries, English, Display

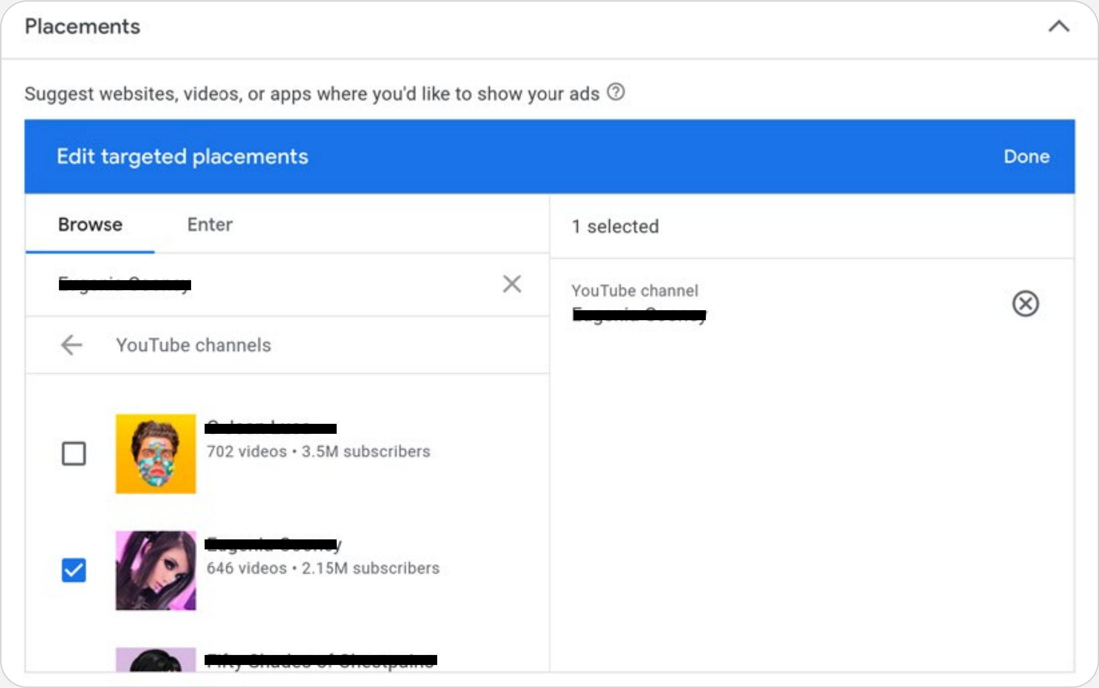
Weekly impressions
–

Gender

Age

Parental status

Figure 14: Google's audience builder, showing how end-users can be targeted based on mental health and eating disorder apps they have downloaded



Placements

Suggest websites, videos, or apps where you'd like to show your ads

Edit targeted placements | Done

Browse | Enter

[redacted] ×

← YouTube channels

□ [redacted]
702 videos • 3.5M subscribers

☑ [redacted]
646 videos • 2.15M subscribers

[redacted]

1 selected

YouTube channel
[redacted]

Figure 15: Google's audience builder, showing how end-users can be targeted based on a popular YouTube channel associated with eating disorders

These findings highlight a systemic issue

These findings are in keeping with existing research, which has highlighted the worrying ability to target unsafe ads at vulnerable demographics on Meta.[26] Meta's ad manager system was found to allow advertisers to target children they had profiled in 'vulnerable' categories such as 13-17 year olds interested in 'Gambling', 'Alcohol', 'Extreme weight loss', 'E cigarettes', etc. More alarmingly, ads were approved to run to each of these vulnerable profiles containing content that posed unique harms, such as ads calling for beach body-ready looks to children interested in extreme weight loss, or ads containing recipes for cocktails made from booze stolen from your parents' liquor cabinet to children interested in alcohol.[27] Further, research has shown how extensive 'vulnerable' advertising profiles are across Australia, with profiles being created that allow for the targeting of 'heavy gamblers', 'problematic alcohol users', families in 'financial distress' and children and young people based on their geolocation.[28]

What this shows

Ad management systems routinely create vulnerabilities in Australian end-users. Placing obligations on platforms to ensure their ad management systems minimise risk would be helpful in creating a comprehensive regulatory framework that protects users. The proposed reforms to the Basic Online Safety Expectations do not include proposals to ensure basic safety obligations regarding ad management systems are in place. We strongly encourage the inclusion of ad management systems in the Expectations.

Placing safety obligations on ad management systems is a necessary but insufficient step. These obligations need to be reinforced by reforms in the *Privacy Act*, where the data collection practices associated with targeted advertising could be prohibited in the first instance. The extent of the privacy violations inherent in this practice are egregious. This research highlights the widespread processing of personal data that reveals sensitive mental health information and help-seeking behaviours to an indiscriminate audience of advertisers, without meaningful consent. This is not fair nor reasonable processing. Turning off this data processing is more 'upstream' and comprehensive than prohibiting the use of this data processing to target advertising in the first instance.

RECOMMENDATIONS

To ensure Australian users remain safe online, Australia's regulatory framework needs to place safety obligations on all systems and elements. As these four examples highlight, multiple functionalities of a platform can create risks for users, and some platforms have reduced risks on some of their systems more than others. *All* platforms should be designing *all* of these systems and elements in ways that actively mitigate risks. That platforms have not done this of their own accord highlights the failings of a light-touch approach to platform regulation; strong regulation is needed to ensure that platforms proactively identify risks and are accountable for adequately mitigating them.

The *Online Safety Act* will simply not be effective if it becomes a 'Christmas tree bill', based on an endless list of systems (or indeed designated risks). Relying *exclusively* on downstream designation requires constant amendments and exacerbates the issue of regulatory lag as new systems are generated and legislation scrambles to patch new gaps. Harms to the public will routinely occur in these intervening periods, which will be outpaced by rapidly evolving technologies. The regulatory framework needs to be comprehensive, upstream, and future-proofed. This is in keeping with emerging international norms. Figure 16 sets out different approaches to identifying systems on digital platforms for safety oversight.

Systems ‘identified’ in the DSA as subject to risk assessment criteria for Very Large Online Platforms	Systems ‘identified’ in the UK OSA as requiring measures to ensure duties of care are met across Platforms	Systems ‘identified’ in the proposed BOSE as being subject to expectations regarding reasonable steps
Recital 84 outlines that services should “focus on the systems or other elements that may contribute to the risks” and lists a number of examples. Other systems and elements specifically listed across the legislation include: 1. Recommender systems 2. ‘Safety by design’ settings for minors 3. Dark patterns and design of interfaces 4. Advertising systems 5. Content moderation systems 6. Notice action and complaint mechanisms 7. Trusted flagger systems 8. Terms and conditions	The duties of care laid out in the Act “apply across all areas of a service, including the way it is designed, operated and used as well as content present on the service” and lists the following areas as requiring measures: 1. Regulatory compliance and risk management arrangements 2. Design of functionalities, algorithms and other features 3. Policies on terms of use; 4. Policies on user access to the service or to particular content present on the service, including blocking users from accessing the service or particular content 5. Content moderation, including taking down content 6. Functionalities allowing users to control the content they encounter 7. User support measures 8. Staff policies and practices	1. Generative AI capabilities 2. Recommender systems 3. User controls 4. ‘Safety by design’ settings for minors (via best interests proposal in subsection 6(2)(A)) 5. Enforcement of terms of use (14(1A)) 6. Complaints and reporting systems (14(3)) We note that some aspects of staff practices covered by the UK’s OSA may be addressed by proposals to amend paragraph 6(3)(f), to add in a suggested example that services assessing whether business decisions will have a significant adverse impact on the ability of end-users to use the service in a safe manner. Further, elements of the DSA’s requirements around terms and conditions regarding understandability are being explored in the *Privacy Act Review*.

Figure 16: A non-exhaustive list of systems and elements ‘identified’ in various safety legislation and proposals, by jurisdiction

We recommend combining an approach that focuses on all risks with a regulatory framework that meaningfully shift the risk mitigation (or safety assurance) burden squarely onto platforms. One approach is via a duty of care model. Under existing legal frameworks, a duty of care approach places a duty on the people who control and are responsible for the hazardous environment. This approach is applicable to the online environment.[29] An overarching duty accompanied by an indicative but expressly non-exhaustive list of associated systems and elements creates both the necessary guidance for protective coverage while also placing an appropriate burden on platforms to ensure safety preventatively and proactively.

There is strong public support for regulations ensuring that platforms take basic safety steps across a full range of systems and processes. In January 2024, Reset.Tech commissioned YouGov to poll 1,005 Australian adults. We found overwhelming support for including expectations regarding more systems—such as advertising systems and content moderation systems—and all systems in general (see Figure 17).

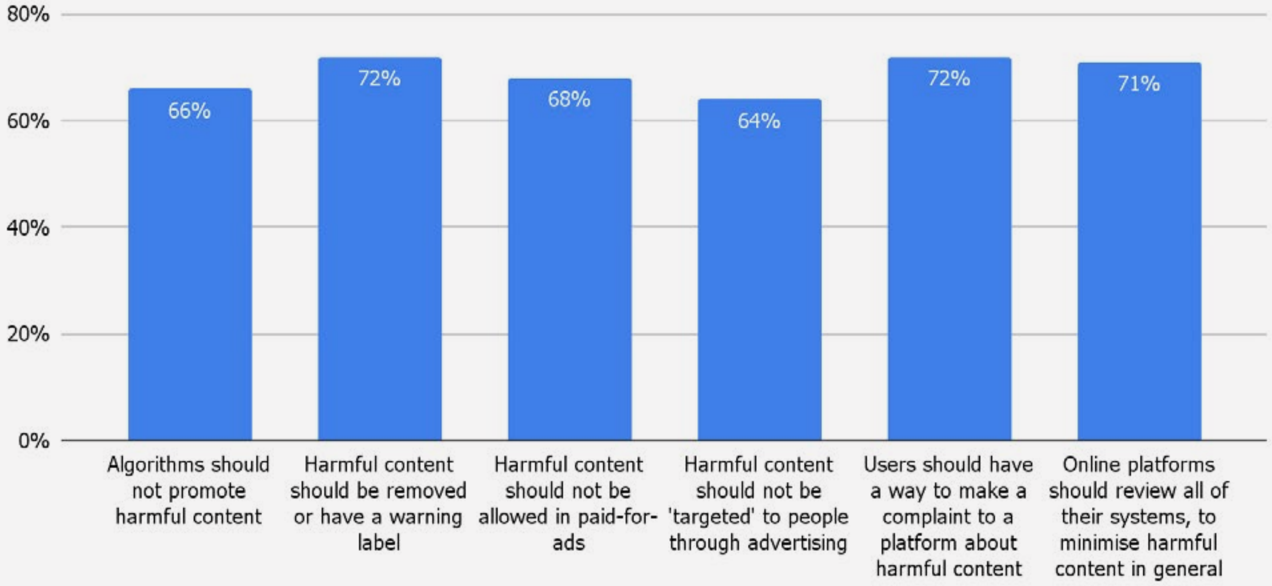


Figure 17: Responses to the question 'which of the following do you think online safety regulations should require?' (n=1,005)

There is strong public support for addressing all online risks in the *Online Safety Act* (see Figure 18). There are too many to enumerate, but Australians support broad coverage for the Act. An overarching duty of care could ensure that all relevant risks and issues are addressed.

Photo credit: Jamillah Knowles & Reset.Tech Australia / Better Images of AI / People with phones / CC-BY 4.0

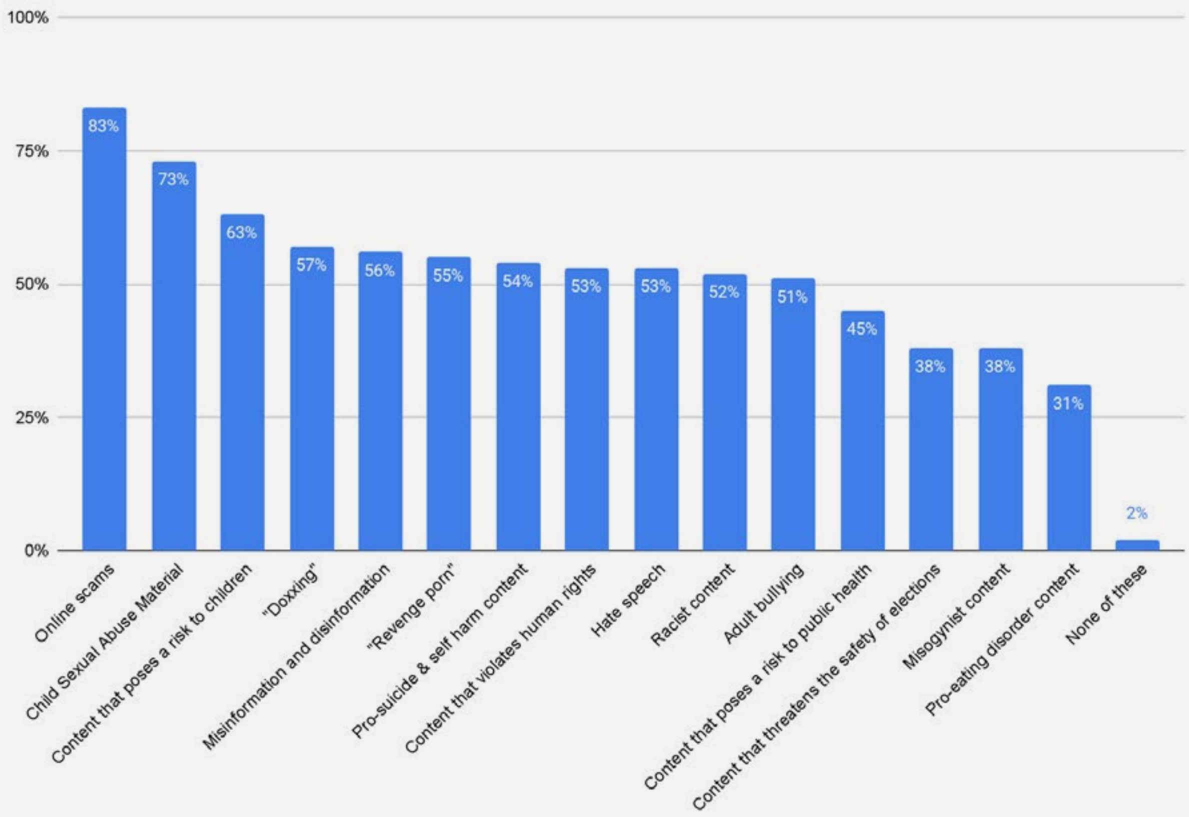
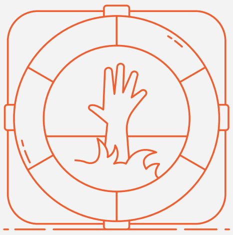


Figure 18: Responses to the question 'which of the following would you like Australia's online safety regulation to address?' (n=1,005)



Our recommendations are as follows:

- The Basic Online Safety Expectations should be amended to include additional expectations that online service providers take reasonable steps regarding all systems and elements involved in the operation of their service. Designating specific systems is inevitably going to create gaps in protection – in systems such as content moderation systems, ad approval systems or ad moderation systems as identified in this report, or in others we have not investigated such as search systems or user-engagement systems – so safety expectations should be extended to all systems and elements. These expectations could require platforms to identify risk-creating systems and elements and consider end-user safety in the design, implementation and maintenance of these systems.
- The *Online Safety Act* Review should implement:
  - An overarching duty of care on platforms to ensure they protect end-users from reasonably foreseeable risks across all their systems and elements
  - Risk assessments and risk mitigation obligations, to ensure the duty of care is realised across all systems and elements
  - Meaningful transparency measures to make visible the risks and mitigation measures created by systems and elements, including sharing risk assessments with regulators, annual public transparency reports with predetermined requirements about the nature of the information that must be shared, independent audits or risk assessments and transparency reports, researcher access to public interest data and a robust ad repository
  - Strong accountability and enforcement mechanisms, including the ability for regulators to compel redress, enhanced civil penalties and the ability to ‘turn off’ services demonstrating persistent failures in cases where all other responses have been exhausted.

Report

ENDNOTES

[1] Recital 84, EU 2022 *Digital Services Act* https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R2065

[2] UK 2023 *Online Safety Act* https://www.legislation.gov.uk/ukpga/2023/50/enacted

[3] National Eating nd Disorders Collaboration nd *Eating Disorders in Australia* https://nedc.com.au/eating-disorders/eating-disorders-explained/eating-disorders-in-australia, see also see also Phillipa Hay, Phillip Aouad, Anvi Le, Peta Marks, Danielle Maloney, Stephen Touyz, & Sarah Maguire 2023 'Epidemiology of eating disorders: population, prevalence, disease burden and quality of life informing public policy in Australia—a rapid review' *Journal of Eating Disorders* https://doi.org/10.1186/s40337-023-00738-7

[4] Patricia Westmoreland, Mori Krantz & Philip Mehler, 2016 'Medical complications of anorexia nervosa and bulimia' *The American Journal of Medicine* https://doi.org/10.1016/j.amjmed.2015.06.031

[5] Marika Tiggemann & Amy Slater 2017 'Facebook and body image concern in adolescent girls: A prospective study' *International Journal of Eating Disorders* http://dx.doi.org/10.1002/eat.22640

[6] Rachel Cohen, Toby Newton-John & Amy Slater 2017 'The relationship between Facebook and Instagram appearance focused activities and body image concerns in young women' *Body Image* http://dx.doi.org/10.1016/j.bodyim.2017.10.002,

[7] Simon Wilksch, Anne O'Shea, Phoebe Ho, Sue Byrne & Tracy Wade 2020 'The relationship between social media use and disordered eating in young adolescents' *International Journal of Eating Disorders* https://doi.org/10.1002/eat.23198

[8] Paula Padín, Rubén González-Rodríguez, Carmen Verde-Diego & Raquel Vázquez-Pérez 2021 'Social media and eating disorder psychopathology: A systematic review' *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* https://doi.org/10.5817/CP2021-3-6

[9] Ben Popkin 2018 'As algorithms take over, YouTube's recommendations highlight a human problem' *NBC News* https://www.nbcnews.com/tech/social-media/algorithms-take-over-youtube-s-recommendations-highlight-human-problem-n867596

[10] See for example this reddit thread Reddit 2020 *Videos and comments that I previously liked get unliked automatically* https://www.reddit.com/r/Tiktokhelp/comments/gtc1k0/videos_and_comments_that_i_previously_liked_get/

[11] See for example a blog by Real Socialz 2023 *Why are my likes disappearing on TikTok* https://realsocialz.com/why-are-my-likes-disappearing-on-tiktok/

[12] Wall Street Journal 2021 'How TikTok's Algorithm Figures Out Your Deepest Desires' *Wall Street Journal* https://www.wsj.com/video/series/inside-tiktoks-highly-secretive-algorithm/investigation-how-tiktok-algorithm-figures-out-your-deepest-desires/6C0C2040-FF25-4827-8528-2BD6612E3796

[13] Reset.Tech 2022 *Designing for Disorder* https://au.reset.tech/news/designing-for-disorder-instagram-s-pro-eating-disorder-bubble-in-australia/

[14] Ralph Housego & Rys Farthing 2022 'Social Grooming' *AQ Magazine* https://www.jstor.org/stable/27161413

[15] Reset.Tech & IDS 2022 *Algorithms as a weapon against women: How YouTube lures boys and young men into the 'Manosphere'* https://au.reset.tech/news/algorithms-as-a-weapon-against-women-how-youtube-lures-boys-and-young-men-into-the-manosphere/

[16] Reset.Tech 2022 *Designing for Disorder* https://au.reset.tech/news/designing-for-disorder-instagram-s-pro-eating-disorder-bubble-in-australia/

[17] See for example, Australian Child Rights Taskforce 2023 *Letter to the eSafety Commissioner* https://childrightstaskforce.org.au/wp-content/uploads/2023/01/Online-Safety-Codes_-ACRT-letter-to-eSafety.pdf

[18] Reset.Tech Australia 2021 *An investigation into TikTok's data processing practices* https://au.reset.tech/news/surveilling-young-people-online-an-investigation-into-tiktok-s-data-processing-practices/

[19] See our research series Reset.Tech 2023 *Risks to Minors* https://www.reset.tech/resources/risktominors/

[20] For a description of TikTok's, Facebook's and X's ad approval systems see Reset.Tech 2023 *How do platforms handle electoral misinformation in paid-for advertising? An experimental evaluation using the Voice referendum* https://au.reset.tech/news/report-misinformation-in-paid-for-advertising/

[21] See for example, Meta 2021 Breaking Down Facebooks *Ad Review Process* https://www.facebook.com/business/news/facebook-ad-policy-process-and-review

[22] Dylan Williams 2020 *How Facebook lets you break Australian electoral laws in under 15 minutes* https://medium.com/ausreset/how-facebook-lets-you-break-australian-electoral-laws-in-under-15-minutes-7db5619ccc9b

[23] Reset.Tech 2021 *Profiling Children for Advertising* https://au.reset.tech/news/profiling-children-for-advertising-facebooks-monetisation-of-young-peoples-personal-data/

[24] Tech Transparency Project 2021 *Facebook's Repeat Fail on Harmful Teen Ads* https://www.techtransparencyproject.org/articles/facebooks-repeat-fail-harmful-teen-ads

[25] Reset.Tech Australia 2023 *Misinformation in paid-for advertising* https://au.reset.tech/news/report-misinformation-in-paid-for-advertising/

[26] Reset.Tech Australia 2020 *Profiling Children for Advertising* https://au.reset.tech/news/profiling-children-for-advertising-facebooks-monetisation-of-young-peoples-personal-data/

[27] This research was subsequently repeated in the US, see Tech Transparency Project 2021 *Pills, Cocktails, and Anorexia: Facebook Allows Harmful Ads to Target Teens* https://www.techtransparencyproject.org/articles/pills-cocktails-and-anorexia-facebook-allows-harmful-ads-target-teens

[28] Reset.Tech 2023 *Australians for Sale: Targeted Advertising, Data Brokering and Consumer Manipulation* https://au.reset.tech/news/coming-soon-australians-for-sale-report/

[29] See for example, Lorna Woods 2019 'Online harm reduction – a statutory duty of care and regulator' *Carnegie Foundation* http://dx.doi.org/10.2139ssrn.4003986

APPENDICES

FOR

NOT JUST ALGORITHMS: ASSURING USER SAFETY ONLINE WITH SYSTEMIC REGULATORY FRAMEWORKS

Reset.Tech Australia,
with support from Dr Hannah Jarman

March 2024

POLICY REPORT

Reset.
AUSTRALIA

APPENDIX 1: PLATFORM POLICIES AROUND PRO-EATING DISORDER CONTENT

TikTok's community guidelines

TikTok's community guidelines outline that the platform "*Remove(s) violative content from the platform that breaks our rules.*"[1] Specifically, when it comes to pro-restrictive eating disorder material, they state: *"We do not allow showing or promoting disordered eating or any dangerous weight loss behaviors."*[2] They describe these as:

- *"Disordered eating includes extreme dieting or fasting, bingeing, and intentional vomiting.*
- *Dangerous weight loss behaviors include compulsive exercise, and using potentially harmful medication or supplements."*

They specifically note that they do not allow:

- *"Showing, promoting, or requesting coaching for disordered eating and other dangerous weight loss behaviors.*
- *Showing or describing extremely low-calorie daily food consumption, and diets associated with disordered eating.*
- *Showing or promoting unhealthy body measurement and "body checking" trends, such as comparing body part size to household objects."*

They note that they do allow *"showing or describing fitness routines and nutrition that are not primarily focused on extreme weight loss, such as preparing for competitive sports, marathon training, and body building competitions."*

According to their guidelines, TikTok should remove violative content when they become aware of it.

Instagram's community guidelines

Instagram's community guidelines[3] outline that the platform aims to:

> *"Maintain [a] supportive environment by not glorifying self-injury. The Instagram community cares for each other, and is often a place where people facing difficult issues such as eating disorders, cutting or other kinds of self-injury come together to create awareness or find support. … Encouraging or urging people to embrace self-injury is counter to this environment of support, and we'll remove it or disable accounts if it's reported to us. We may also remove content identifying victims or survivors of self-injury if the content targets them for attack or humour."*

Instagram describes self-injury using Meta's head terms:[4]

> *"While we do not allow people to intentionally or unintentionally celebrate or promote suicide or self-injury, we do allow people to discuss these topics because we want Facebook to be a space where people can share their experiences, raise awareness about these issues, and seek support from one another.*
>
> *We define self-injury as the intentional and direct injuring of the body, including self-mutilation and eating disorders. We remove any content that encourages suicide or self-injury, including fictional content such as memes or illustrations and any self-injury content that is graphic, regardless of context.*

[Do not post]

- *Content that focuses on depiction of ribs, collar bones, thigh gaps, hips, concave stomach or protruding spine or scapula when shared together with terms associated with eating disorder*
- *Content that contains instructions for drastic and unhealthy weight loss when shared together with terms associated with eating disorders."*

According to their guidelines, Instagram should remove violative content when they become aware of it.

X's community guidelines

Xs community guidelines[5] state that they prohibit content that promotes or encourages self-harm behaviours.

> *"...you can't promote, or otherwise encourage, suicide or self-harm. We define promotion and encouragement to include statements such as "the most effective", "the easiest", "the best", "the most successful", "you should", "why don't you". Violations of this policy can occur via Posts, images or videos, including live video.*

We define suicide to be the act of taking one's own life. We define self-harm to include:

- *self-inflicted physical injuries e.g., cutting; and*
- *eating disorders e.g., bulimia, anorexia.*

Violations of this policy include, but are not limited to:

- encouraging someone to physically harm or kill themselves;
- asking others for encouragement to engage in self-harm or suicide, including seeking partners for group suicides or suicide games; and
- *sharing information, strategies, methods or instructions that would assist people to engage in self-harm and suicide."*

They do not provide any relevant examples of what might be considered promoting or encouraging eating disorders such as bulimia and anorexia.

According to their guidelines, X should remove violative content when they become aware of it.

APPENDIX 2: CODEBOOK OF PRO-EATING DISORDER CONTENT

Our content moderation experiment and recommender system experiment required accurately identifying pro-eating disorder content. To do this, we used the Codebook below. This Codebook was initially and substantively developed with support from academics and practitioners from the InsideOut Institute for Eating Disorders (Australia's National Eating Disorder Research and Clinical Excellence Institute, Sydney University & Sydney Local Health District)[6] and subsequently improved in conversations with Hannah Jarman from Deakin University. We are indebted to their expertise in developing this Codebook.

This Codebook describes content that would violate all major platforms' community guidelines, and is considered harmful by eating disorder experts.

We only include content that is explicitly *pro*-restrictive eating disorder. **We describe pro-restrictive eating disorder content as content that shows, promotes or normalises dangerous weight loss, or dangerous restrictive dieting behaviours, and/or glorifies bodies with extremely low weight,** such as:

- Content that focuses primarily on or highlights bone structures, such as ribs, collar bones, thigh gaps, hips, concave stomach or protruding spine or scapulas; this includes content where people are posing or deliberately exposing body parts to highlight extreme thinness
- Content that primarily depicts thigh gaps, finger gaps, or other signs of extremely low body mass
- Before and after weight loss content, where the after picture shows either bone structures or clinically unhealthy BMI rates
- Content that displays or notes BMI rates below 18.5, or expresses goals for BMI rates below 18.5
- Content regarding weight check ins or Goal Weights or Ultimate Goal Weights that would be associated with an unhealthy BMI (such as content where someone describes their weight as 52 kg, and their height as 5'8")
- Showing or promoting unhealthy body measurement and 'body checking' trends, such as comparing body part size to household objects
- Content that congratulates people for reaching an unhealthy weight, such as a BMI below 18.5 or having visible bones, or that encourages them to aspire to this
- Content that shows or promotes extremely low-calorie daily food consumption (less than 500 calories per meal, or 1,500 per day) when combined with eating disorder terms or diets associated with disordered eating
- Content that congratulates people for restricting their eating to less than 500 calories per meal, or 1,500 per day or that encourages them to achieve this
- Content that describes having an eating disorder as a positive outcome or depicts eating disorders in a desirable or accessible light (e.g. tweets that say 'restricting is easy, will power lets me just eat water') or encourages someone to join them in their restriction/ED (e.g., ana buddy wanted)

We do not want to include content that only talks about eating disorders, content that documents the day-to-day life of people affected by eating disorders, nor content that is about struggling with disorders or roads-to-recovery. We do not include:

- Content that depicts bone structures, thigh gaps or BMIs in association with text or images that describe wanting to recover or gain weight
- Content that depicts bone structures, thigh gaps or BMIs where associated content (terms etc.) indicated that the person in the picture was trying to put on weight or otherwise documenting an attempt at recovery
- Content that just features extremely skinny people, who may or may not be affected by restrictive eating disorders, who are just documenting their lives (such as playing guitar, on a walk), where the content does not explicitly centre around their weight or include associated terms; this does not include images where people are deliberately posing and focusing on their visible bone structures, or thigh gaps etc
- Recovery diaries or recovery stories
- Content that talks about the difficulties of having a restrictive eating disorder, or talks about day-to-day issues
- Content that depicts bone structures, thigh gaps or BMIs in a medical or humanitarian context (e.g. documenting a famine or person ill from non-eating disorder diseases)
- Content where it is unclear if it is referencing eating disorders (e.g. memes about going to the fridge, losing willpower, and eating 1,000 calories, where it was unclear from the meme if that was all they ate during the day or just a big ‘snack’ they regret)
- Low calorie diet content that does not include eating disorder terms, such as for content associated with ‘diabetes friendly’ diets, or general weight loss diets
- Images of professional athletes, such as ultra marathon runners or ballerinas
- Exercise ‘for weight loss’ content

APPENDIX 3: PLATFORM POLICIES AROUND PRO-EATING DISORDER CONTENT IN PAID-FOR ADVERTISING

Platforms have policies prohibiting certain types of harmful content, including pro-eating disorder content, from being featured in paid-for ads.

TikTok's advertising policies

TikTok's advertising policies prohibits:

> *"Ads promoting weight loss or management fasting products or services. Ads promoting weight loss/management supplements (including but not limited to fat-burning pills, appetite suppressants, weight loss teas, or lollipops)."*[7]

Instagram's advertising policies

Instagram's (Meta's) advertising policies state that:

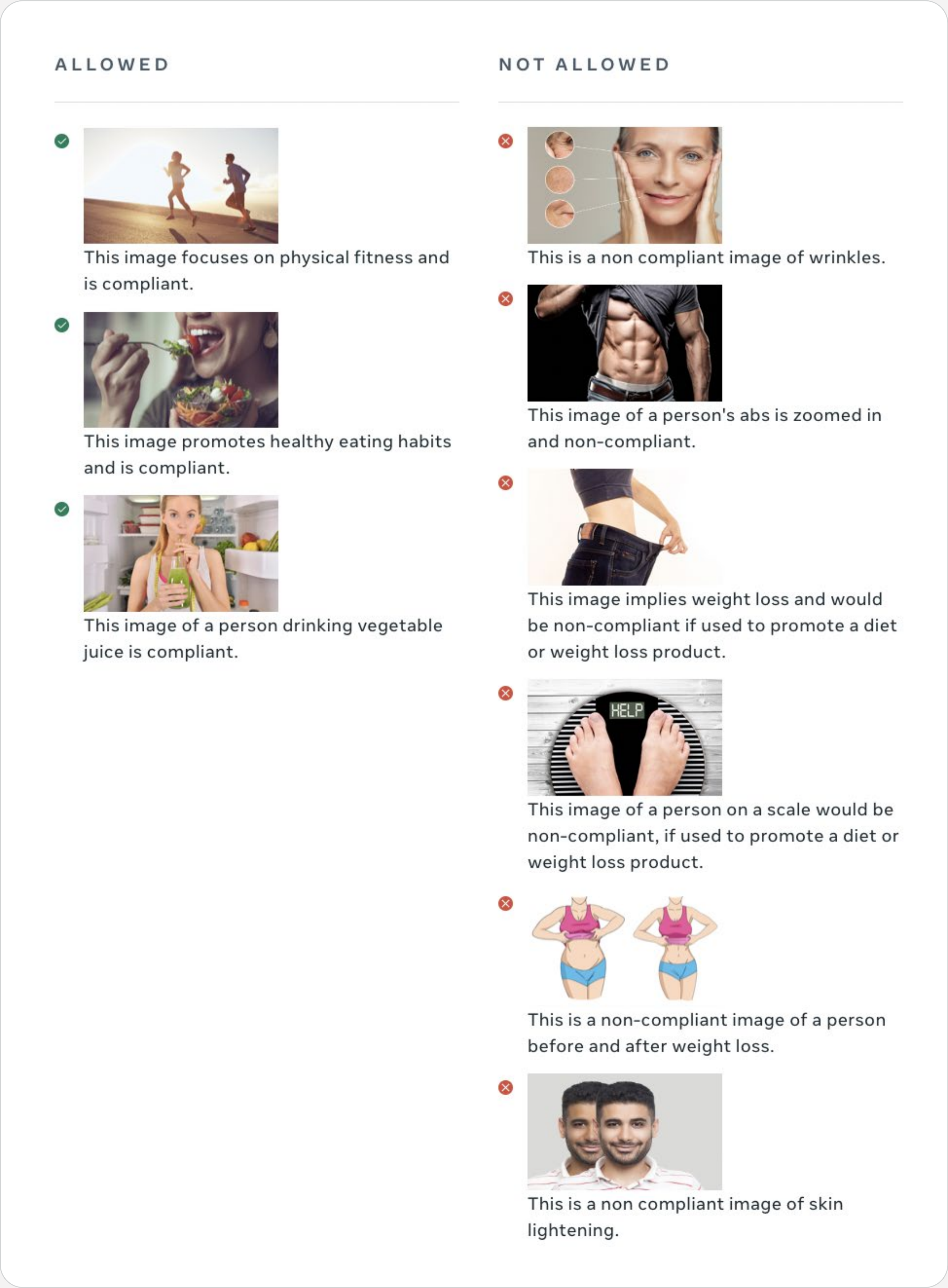
> *"Ad content must not imply or attempt to generate negative self-perception in order to promote diet, weight loss or other health-related products.*

Ads can't:

- *Declare or imply there is a perfect body type or appearance that one should aspire to*
- *Promote or reinforce negative or unhealthy body images*
- *Exploit insecurities to conform to certain beauty standards*
- *Contain distasteful messaging that could make people feel negatively about the way they look*
- *Promote an unhealthy relationship with food or exercise*
- *Show close-up imagery on the health condition of a person*
- *Feature body shaming of any type*

Note: Ads for cosmetics, hair extensions, other similar cosmetic or non-permanent beauty products or digital editing apps aren't within scope of this policy."[8]

This is further explained with examples (see Figure 19).



ALLOWED

- ✓ This image focuses on physical fitness and is compliant.
- ✓ This image promotes healthy eating habits and is compliant.
- ✓ This image of a person drinking vegetable juice is compliant.

NOT ALLOWED

- ✗ This is a non compliant image of wrinkles.
- ✗ This image of a person's abs is zoomed in and non-compliant.
- ✗ This image implies weight loss and would be non-compliant if used to promote a diet or weight loss product.
- ✗ This image of a person on a scale would be non-compliant, if used to promote a diet or weight loss product.
- ✗ This is a non-compliant image of a person before and after weight loss.
- ✗ This is a non compliant image of skin lightening.

Figure 19: Visual guidance regarding paid-advertising policies on Meta[9]

X’s advertising policies

X’s advertising policies state that:

> *“When advertisers on X choose to promote their content with X Ads, their account and content become subject to an approval process. The approval process is designed to support the quality and safety of the X Ads platform. This process helps X check that advertisers are complying with our advertising policies.*
>
> *In addition to X Ads Policies, advertisers must follow X’s Terms of Service, X Rules, and all the policies on our Help Center governing use of our services.”*[10]

X claims that it prohibits ads that contain inappropriate content globally. While all content on X is subject to the ‘X Rules’, additional restrictions are placed on advertising content. This includes:

“Harmful Weight Loss Content

- Ads must not promote weight loss content that impacts physical and mental health and body image. Examples:
  - Content that is reasonably considered to ‘body-shame’ the customer
  - Content that encourages, glamorizes, or promotes unhealthy or unsafe eating behaviors or eating disorders

Further, X prohibits knowingly marketing or advertising a list of harmful products to minors. This includes:

- *Weight loss products and services and content focused on weight loss*
- *Health and wellness supplements (including, but not limited to, health, dietary, food, nutrition, weight loss, and muscle enhancement substances and supplements).”*[11]

Google’s advertising policies

Google’s advertising policies state that ads that depict the following content are not allowed on their platform:

“Content that threatens or advocates for physical or mental harm on oneself or others. Examples (non-exhaustive): Content advocating suicide, anorexia, or other self-harm…”[12]

APPENDIX 4: PLATFORM AD MANAGEMENT SYSTEMS AND ABILITIES TO TARGETED END-USERS INTERESTED IN PRO-EATING DISORDER CONTENT

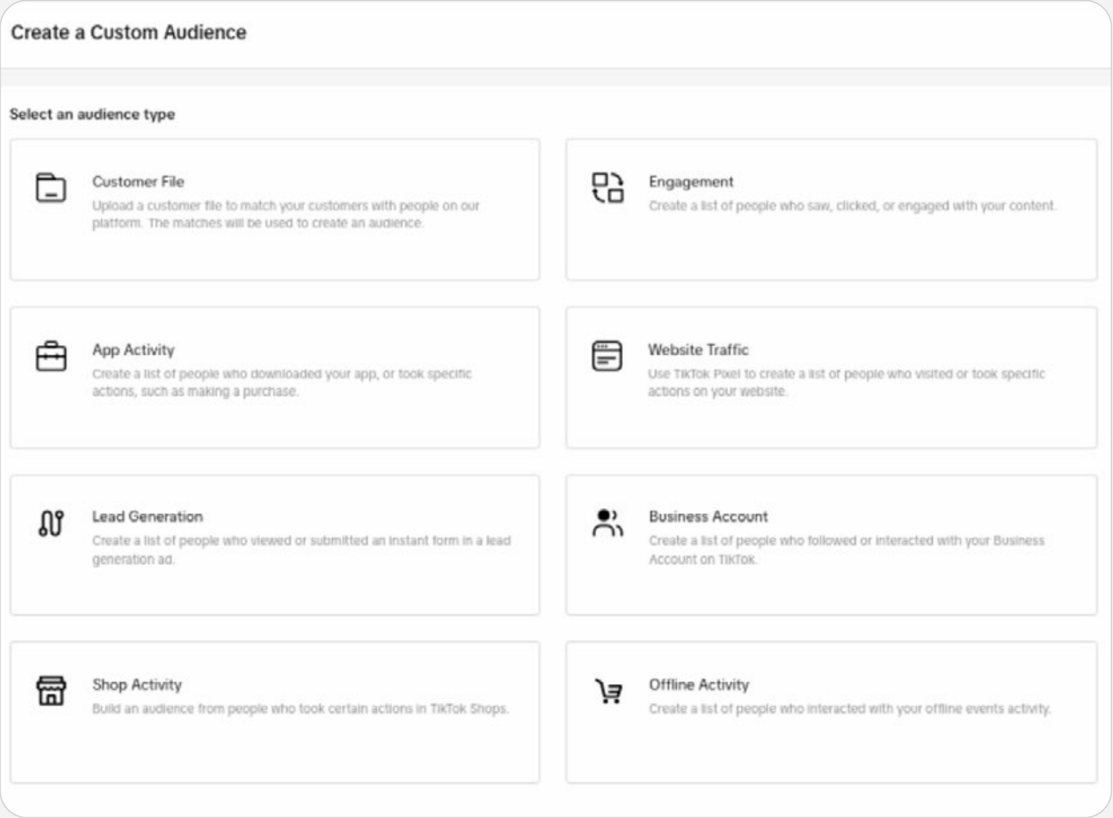
Platforms' ad management systems process end-users' data in multiple ways to enable the targeting of users who may be interested in pro-eating disorder content.

TikTok's ad management system

Advertisers could target end-users interested in pro-eating disorder content by using: engagements on TikTok, enhanced by including 'lookalike' audiences; off-app data such as customer files, website traffic or app downloads and enhanced by including 'lookalike' audiences, and; to some extent hashtag targeting. These processes would not be simple or straight forward, but could be very powerful in accuracy.

TikTok's ad management system functions very differently to older online platforms. TikTok is heavily reliant on powerful algorithms, and deploys these algorithms to create 'custom audiences' for advertisers. TikTok allows advertisers to create custom audiences (i.e. targets for advertising) based on:

- Customer files: Advertisers can upload their own customer files to 'match' their customers with TikTok users
- Engagement: Advertisers can target end-users who have clicked on or shared content on their TikTok account
- App activity: TikTok tracks some apps that are downloaded on a users' phone and how these apps are used. They allow advertisers to target end-users who have downloaded an advertiser's app onto their phone, or taken a specific action on this app, such as made a purchase
- Website traffic: TikTok tracks websites that users visit and how these websites are used. They allow advertisers to target end-users who have visited specific websites, or have taken a specific action on websites
- Lead generation: Advertisers can target end-users who have viewed or 'clicked on' a lead generation ad posted by the advertiser previously
- Business accounts: Advertisers can target end-users who have interacted with their TikTok account, provided it is a business account
- Shop activity: Advertisers can target end-users who have taken specific actions in TikTok shops
- Offline activity: Advertisers can target end-users who have interacted with their offline events (see Figure 20).



Create a Custom Audience

Select an audience type

Customer File
Upload a customer file to match your customers with people on our platform. The matches will be used to create an audience.

Engagement
Create a list of people who saw, clicked, or engaged with your content.

App Activity
Create a list of people who downloaded your app, or took specific actions, such as making a purchase.

Website Traffic
Use TikTok Pixel to create a list of people who visited or took specific actions on your website.

Lead Generation
Create a list of people who viewed or submitted an instant form in a lead generation ad.

Business Account
Create a list of people who followed or interacted with your Business Account on TikTok.

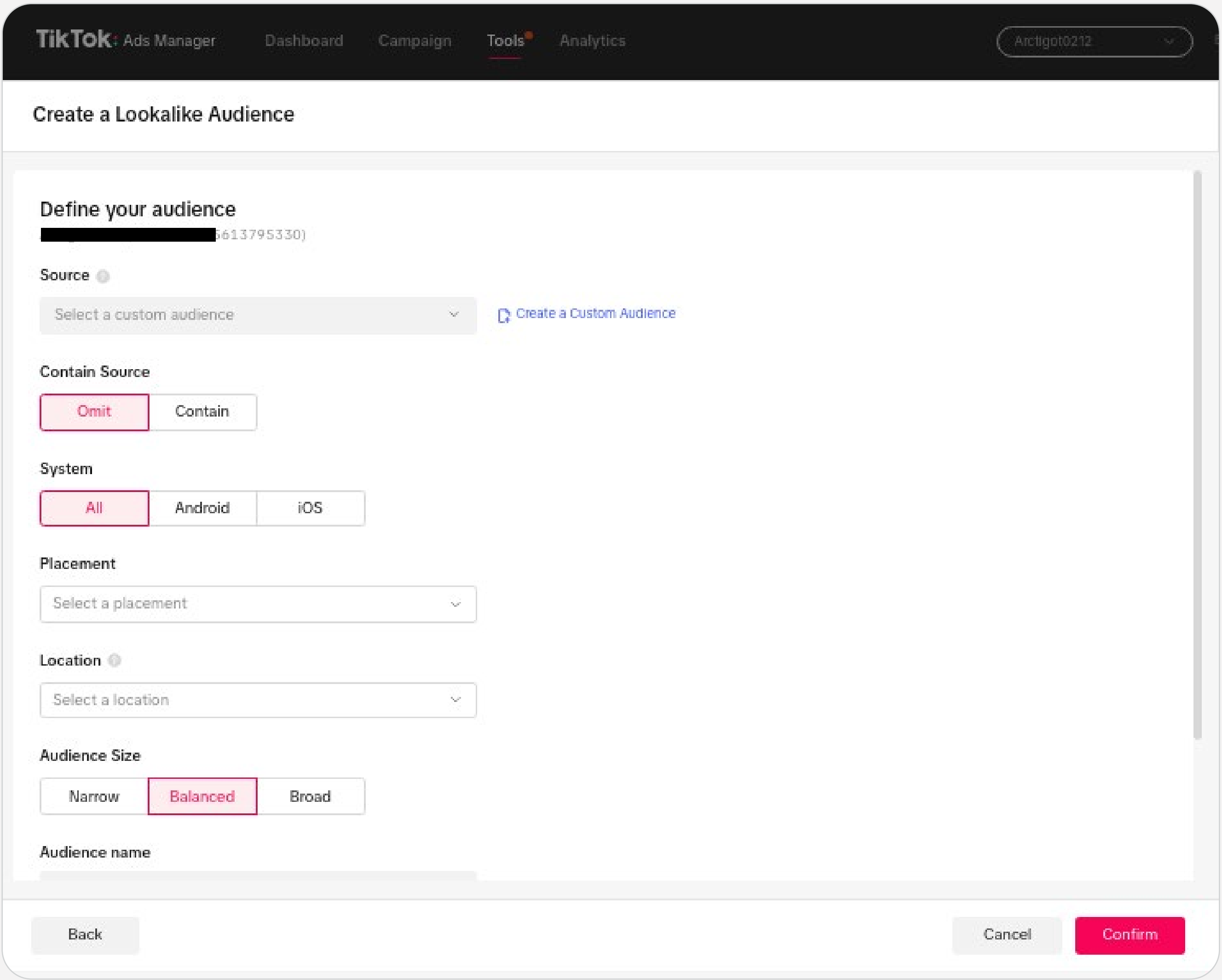
Shop Activity
Build an audience from people who took certain actions in TikTok Shops.

Offline Activity
Create a list of people who interacted with your offline events activity.

Figure 20: TikTok's custom audience builder tool

» *Using engagements on TikTok*

If an advertiser creates and shares pro-eating disorder content, they can then use TikTok's 'engagement tool' to target end-users who have interacted with their content, including pro-eating disorder content. Advertisers can then expand this based on algorithms to target 'lookalike' end-users that have not yet interacted with their content (see Figure 21). We did not test the capacity of this method for ethical reasons, as doing so would require us to post pro-eating disorder content. As far as we are aware, there are no controls based on target groups built from engagements.



Create a Lookalike Audience

Define your audience

[redacted]613795330)

Source: Select a custom audience

Contain Source: [✓ Omit] Contain

System: [✓ All] Android iOS

Placement: Select a placement

Location: Select a location

Audience Size: Narrow [✓ Balanced] Broad

Audience name

Figure 21: TikTok's 'lookalike' audience builder

It is worth noting that traditional measures of transparency, such as 'simple' advertising repositories like the TikTok ad library won't track how these ads are targeted based on engagements (as the targeting is not based on reportable keywords). Further, without keywords for targeting, risky content in ads is almost impossible to detect because it is unsearchable.

» *Using 'third party data' such as customer files, website traffic or app data on TikTok*

Advertisers could also target end-users interested in pro-eating disorder content based on tracking 'off app' data. For example:

- Customer files: If an advertiser had data or had purchased data from a data broker about people interested in pro-eating disorder content,[13] they could upload this as a 'customer file' and target them with ads. This could then be expanded to target 'lookalike' end-users.

  ID data about people who might be interested in pro-eating disorder content is readily available from data brokers. For example, the Xandr file—which documents only the 'customer files' available from one data broker—listed a number of 'lists' available about Australians interested in weight loss. This included numerous lists of Australians such as:

- Those who have visited weight loss centres (Eyeota - APAC Lifesight - Healthcare - Location Visited - Weight Loss and Nutritionist Clinics)
- Those categorised as weight watchers (Eyeota - AU Roy Morgan - Lifestyle - Weight Watcher)
- Those interested in weight loss or dieting (Branded Data > Lifesight > Interest > Weight Loss (BlueKai); International_APAC - Australia Dieting and Weight Loss (Lotame); eXelate Australia Interest - Health - Diet & Weight Loss; eXelate Australia Interest - Health - Weight Loss; Oracle Country-Specific Audiences > Australia (AU) > Hobbies and Interests (Affinity) > Health and Fitness > Wellness > Dieting and Weight Loss (BlueKai))[14]

› Website traffic: If an advertiser had a relevant website or access to the coding of one—such as a weight loss, mental health or eating disorder website—they would be able to target TikTok users who visited or took specific actions on this website. This could then be expanded to include 'lookalike' end-users.

› App data: If an advertiser had a relevant app or access to the coding of one—such as a weight loss, mental health or eating disorder website—they would be able to target TikTok users who have downloaded it or taken specific actions on it. This could then be expanded to include 'lookalike' end-users.

» *Using hashtags on TikTok*

Hashtag targeting is possible on TikTok but is very content-restrictive. Broad hashtags like #diet, #weightloss or #health are allowed but more pro-eating disorder associated hashtags like [redacted] or [redacted] are blocked. It is worth noting that using generic categories to target end-users is probably not the 'usual business practice' for TikTok, as the 'Create a Custom Audience' tool is far more powerful.

Meta's ad management system

Advertisers could target end-users interested in pro-eating disorder content using data derived from Meta sources such as page or account visitor data, enhanced by including 'lookalike' audiences; off-app data such as customer files, website traffic or app downloads and enhanced by including 'lookalike' audiences, and; to a limited extent ad interest data. These processes would not be simple or straightforward, but could be very powerful in accuracy.

Meta allows advertisers to create custom audiences (i.e. targets for advertising) based on:

› Website traffic: Meta tracks websites that users visit and how these websites are used. They allow advertisers to target end-users based on website interactions

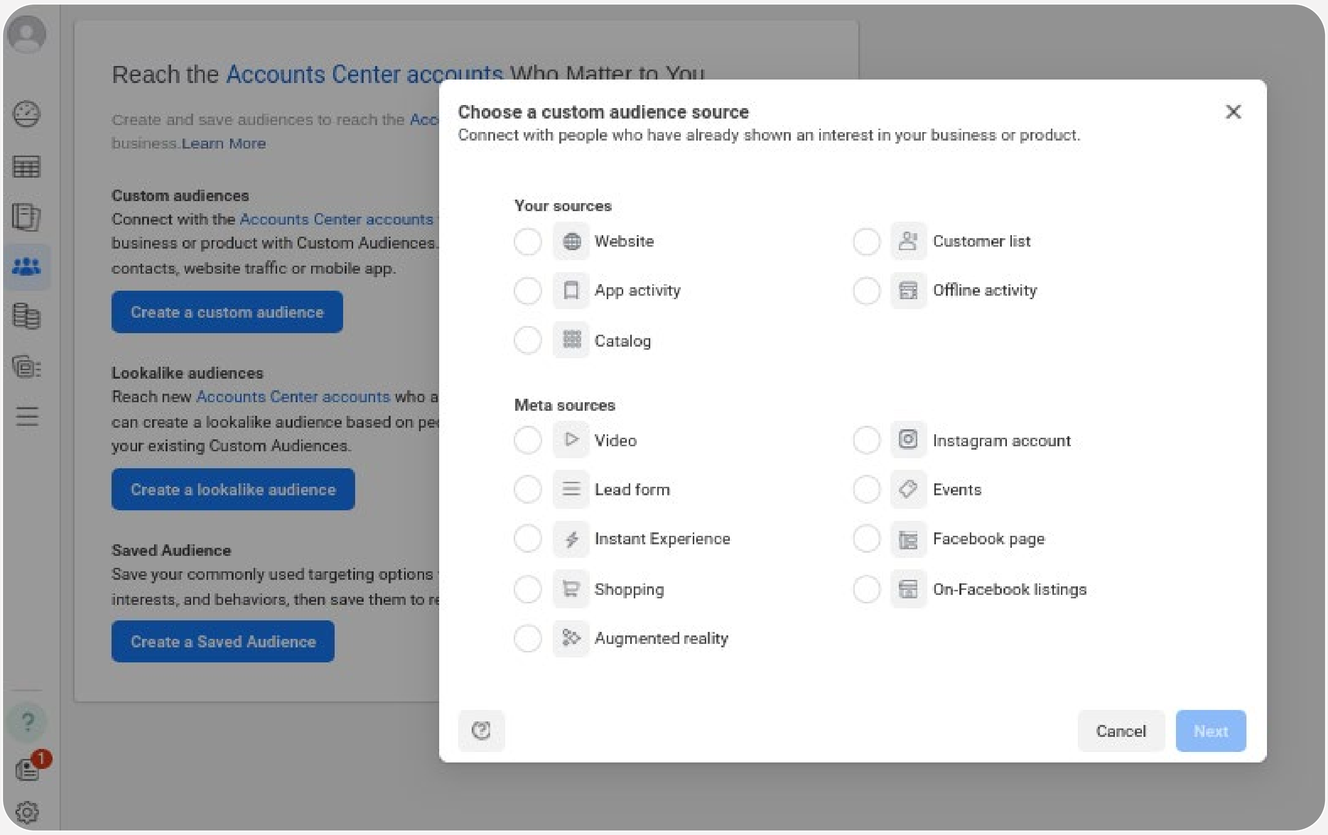
› Customer lists: Advertisers can upload their own customer lists, presumably to 'match' their customers lists with Meta users

› App activity: Meta tracks apps that are downloaded on personal devices and how these apps are used. They allow advertisers to target end-users based on app activity

› Catalogue data: Meta allows advertisers to 'catalogue' data about end-users, which is appears to then allow advertisers to use for targeting[15]

› Offline activity: Advertisers can target end-users based on offline activity

› Data derived from Meta sources; such as user data on Facebook pages, or Instagram accounts (see Figure 22).



Reach the Accounts Center accounts Who Matter to You

Create and save audiences to reach the Acc
business.Learn More

Custom audiences
Connect with the Accounts Center accounts
business or product with Custom Audiences.
contacts, website traffic or mobile app.

Lookalike audiences
Reach new Accounts Center accounts who a
can create a lookalike audience based on pe
your existing Custom Audiences.

Saved Audience
Save your commonly used targeting options
interests, and behaviors, then save them to re

Choose a custom audience source
Connect with people who have already shown an interest in your business or product.

Your sources

○ Website
○ App activity
○ Catalog
○ Customer list
○ Offline activity

Meta sources

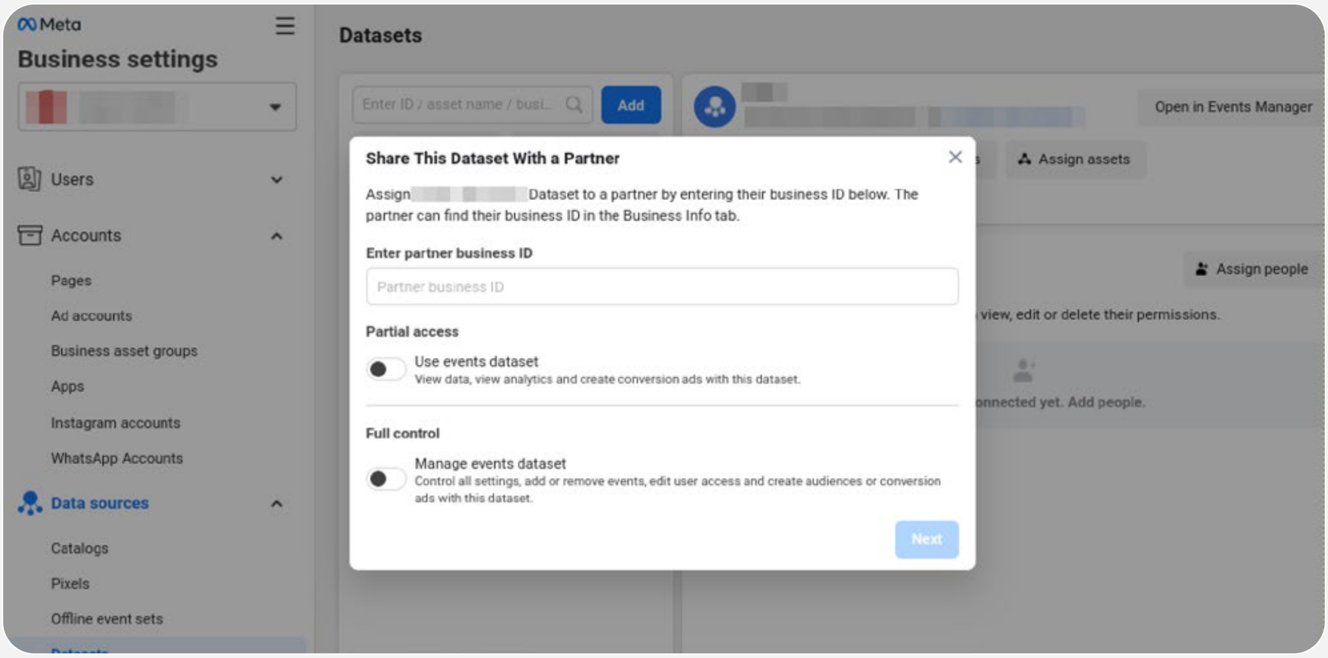
○ Video
○ Lead form
○ Instant Experience
○ Shopping
○ Augmented reality
○ Instagram account
○ Events
○ Facebook page
○ On-Facebook listings

Figure 22: Meta's custom audience builder tool

» *Using data derived from Meta sources such as page or account audience data*

Meta offers business accounts several ways to target end-users who have visited their own Facebook or Instagram pages. This means that if an advertiser manages a pro-eating disorder page or account, they can then target end-users who interact with this content and expand this to lookalikes that don't yet interact with their account using either metadata or 'third-party data' as described below. Advertisers can also share or sell this data with other advertisers on Meta (see Figure 23).

We did not test the capacity of this method for ethical reasons, as we will not post pro-eating disorder content. As far as we are aware, there are no controls based on target groups built from engagements.



Meta

Business settings

[redacted]

Users

Accounts

Pages

Ad accounts

Business asset groups

Apps

Instagram accounts

WhatsApp Accounts

Data sources

Catalogs

Pixels

Offline event sets

Datasets

Enter ID / asset name / busi... Add

[redacted]

Open in Events Manager

Assign assets

Assign people

view, edit or delete their permissions.

onnected yet. Add people.

Share This Dataset With a Partner ✕

Assign [redacted] Dataset to a partner by entering their business ID below. The partner can find their business ID in the Business Info tab.

Enter partner business ID

Partner business ID

Partial access

○ Use events dataset
View data, view analytics and create conversion ads with this dataset.

Full control

○ Manage events dataset
Control all settings, add or remove events, edit user access and create audiences or conversion ads with this dataset.

Next

Figure 23: Meta's advertising management system, showing how end-user data can be shared with other businesses on Meta

» *Using 'third-party data' such as customer files, website traffic or app data on Meta*

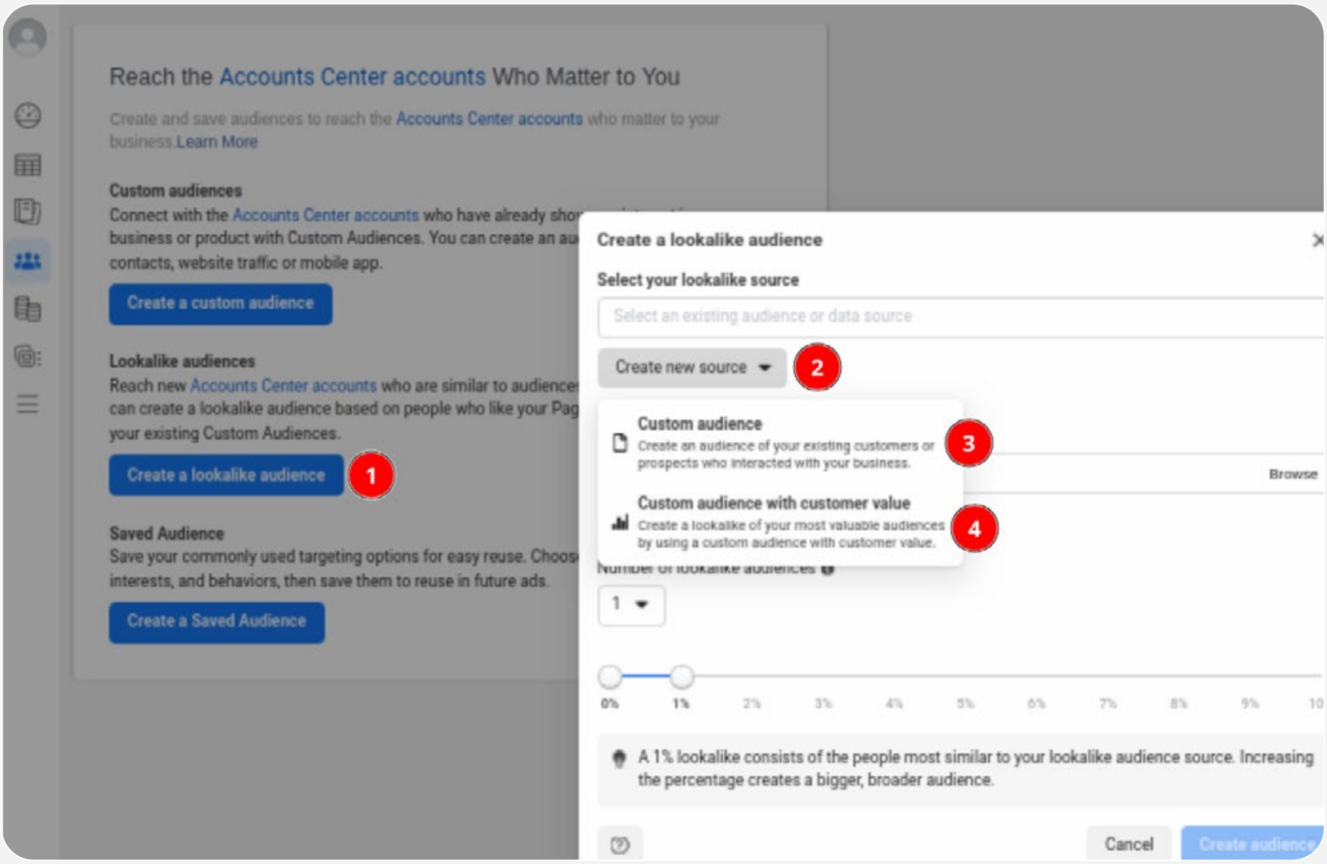
Advertisers could target end-users interested in pro-eating disorder content by tracking 'off Meta' data. For example:

- Customer files: If an advertiser had data or had purchased data from a data broker about people interested in pro-eating disorder content,[16] they could upload this as a 'customer file' and target them with ads. This could then be expanded to target 'lookalike' end-users.

  ID data about people who might be interested in pro-eating disorder content is readily available from data brokers. For example, the Xandr file—which documents only the 'customer files' available from one data broker—listed a number of 'lists' available about Australians interested in weight loss. . This included numerous lists of Australians such as:

  - Those who have visited weight loss centres (Eyeota - APAC Lifesight - Healthcare - Location Visited - Weight Loss and Nutritionist Clinics)
  - Those categorised as weight watchers (Eyeota - AU Roy Morgan - Lifestyle - Weight Watcher)
  - Those interested in weight loss or dieting (Branded Data > Lifesight > Interest > Weight Loss (BlueKai); International_APAC - Australia Dieting and Weight Loss (Lotame); eXelate Australia Interest - Health - Diet & Weight Loss; eXelate Australia Interest - Health - Weight Loss; Oracle Country-Specific Audiences > Australia (AU) > Hobbies and Interests (Affinity) > Health and Fitness > Wellness > Dieting and Weight Loss (BlueKai))[17]

Advertisers also have the ability to target 'lookalike' end-users from data from third party advertising agencies (see Figures 24 and 25). These third party advertising agencies offer a data brokerage service, where advertisers can pay for the ability to use their data without necessarily purchasing the lists themselves. Some of these brokerages offer categories that allow people interested in weight loss to be targeted. As an example, a US company called Versium offers advertisers the ability to mark people in their customer lists as interested in dieting or weight loss (see Figure 26) and allows advertisers to upload Versium audience lists "without effort" (see Figure 27). It is unclear to us how Versium builds its lists or if people consented to this type of data processing.[18] Figure 28 shows a promotional diagram from Versium highlighting the level of detail they report to hold about 'customers' through their products.



Reach the Accounts Center accounts Who Matter to You

Create and save audiences to reach the Accounts Center accounts who matter to your business.Learn More

Custom audiences
Connect with the Accounts Center accounts who have already sho business or product with Custom Audiences. You can create an au contacts, website traffic or mobile app.

Lookalike audiences
Reach new Accounts Center accounts who are similar to audience can create a lookalike audience based on people who like your Pag your existing Custom Audiences.

Saved Audience
Save your commonly used targeting options for easy reuse. Choos interests, and behaviors, then save them to reuse in future ads.

Create a lookalike audience

Select your lookalike source

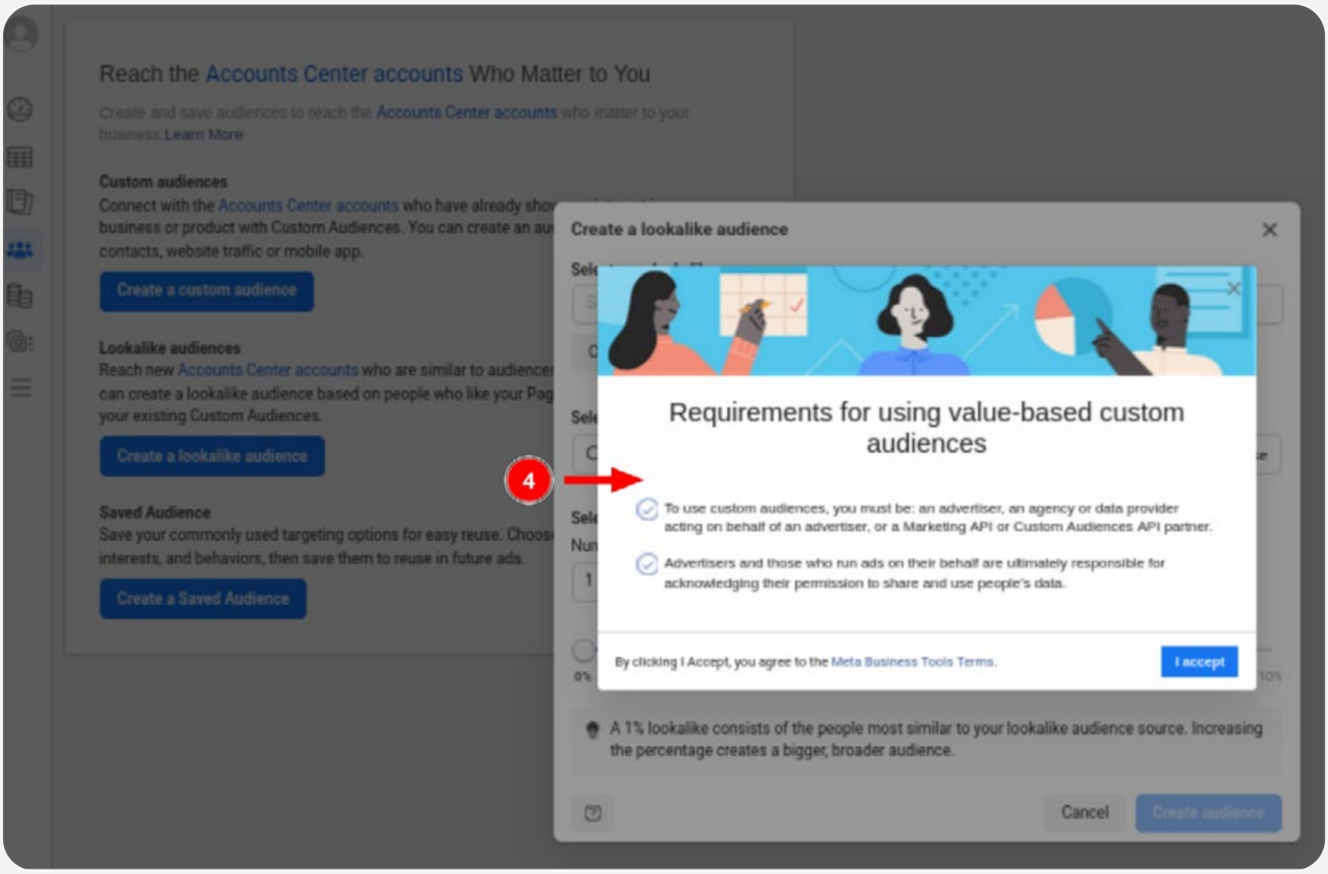
Custom audience
Create an audience of your existing customers or prospects who interacted with your business.

Custom audience with customer value
Create a lookalike of your most valuable audiences by using a custom audience with customer value.

Number of lookalike audiences

A 1% lookalike consists of the people most similar to your lookalike audience source. Increasing the percentage creates a bigger, broader audience.

Figure 24: Meta's options to create 'lookalike' audiences using Meta data and third-party data



Reach the Accounts Center accounts Who Matter to You

Create and save audiences to reach the Accounts Center accounts who matter to your business. Learn More

Custom audiences
Connect with the Accounts Center accounts who have already sho
business or product with Custom Audiences. You can create an au
contacts, website traffic or mobile app.

Create a custom audience

Lookalike audiences
Reach new Accounts Center accounts who are similar to audience
can create a lookalike audience based on people who like your Pag
your existing Custom Audiences.

Create a lookalike audience

Saved Audience
Save your commonly used targeting options for easy reuse. Choos
interests, and behaviors, then save them to reuse in future ads.

Create a Saved Audience

Create a lookalike audience

Requirements for using value-based custom audiences

4

To use custom audiences, you must be: an advertiser, an agency or data provider acting on behalf of an advertiser, or a Marketing API or Custom Audiences API partner.

Advertisers and those who run ads on their behalf are ultimately responsible for acknowledging their permission to share and use people's data.

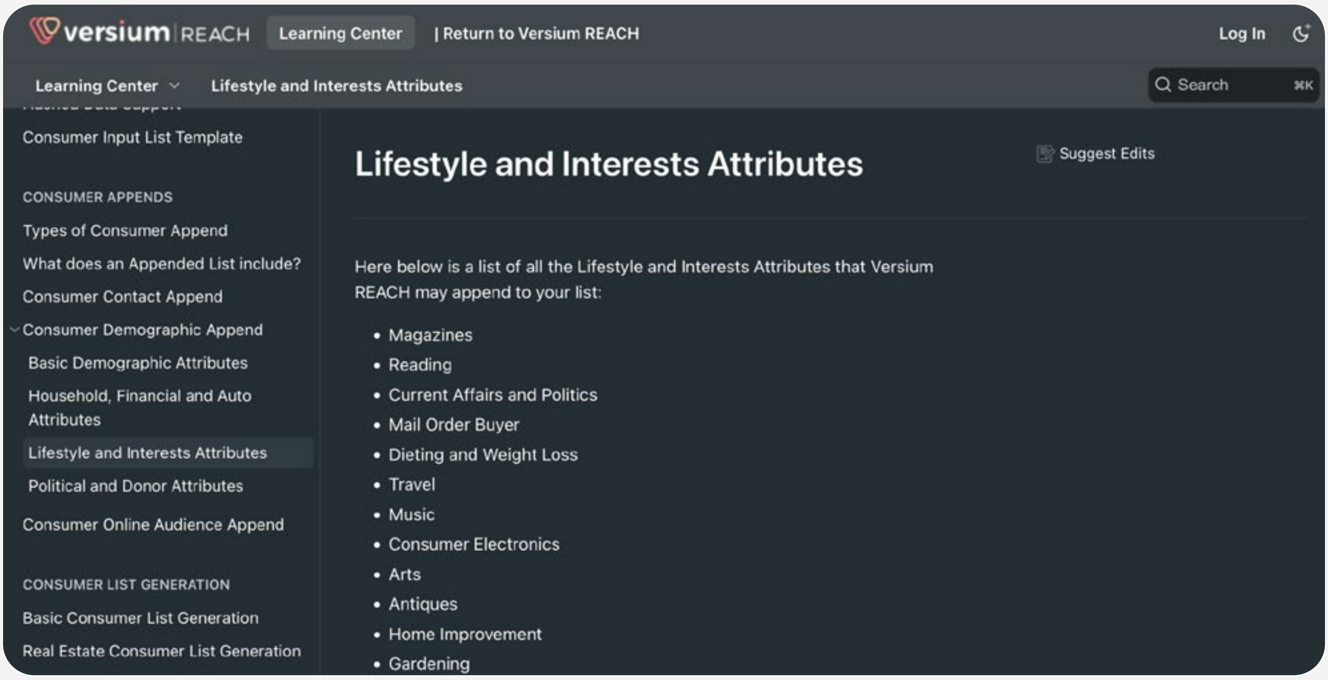
By clicking I Accept, you agree to the Meta Business Tools Terms.

I accept

A 1% lookalike consists of the people most similar to your lookalike audience source. Increasing the percentage creates a bigger, broader audience.

Cancel Create audience

Figure 25: Meta's options to create lookalike audiences using third-party data advertising agency data



versium REACH | Learning Center | Return to Versium REACH | Log In

Learning Center | Lifestyle and Interests Attributes | Search ⌘K

Consumer Input List Template

CONSUMER APPENDS
Types of Consumer Append
What does an Appended List include?
Consumer Contact Append
Consumer Demographic Append
Basic Demographic Attributes
Household, Financial and Auto Attributes
Lifestyle and Interests Attributes
Political and Donor Attributes
Consumer Online Audience Append

CONSUMER LIST GENERATION
Basic Consumer List Generation
Real Estate Consumer List Generation

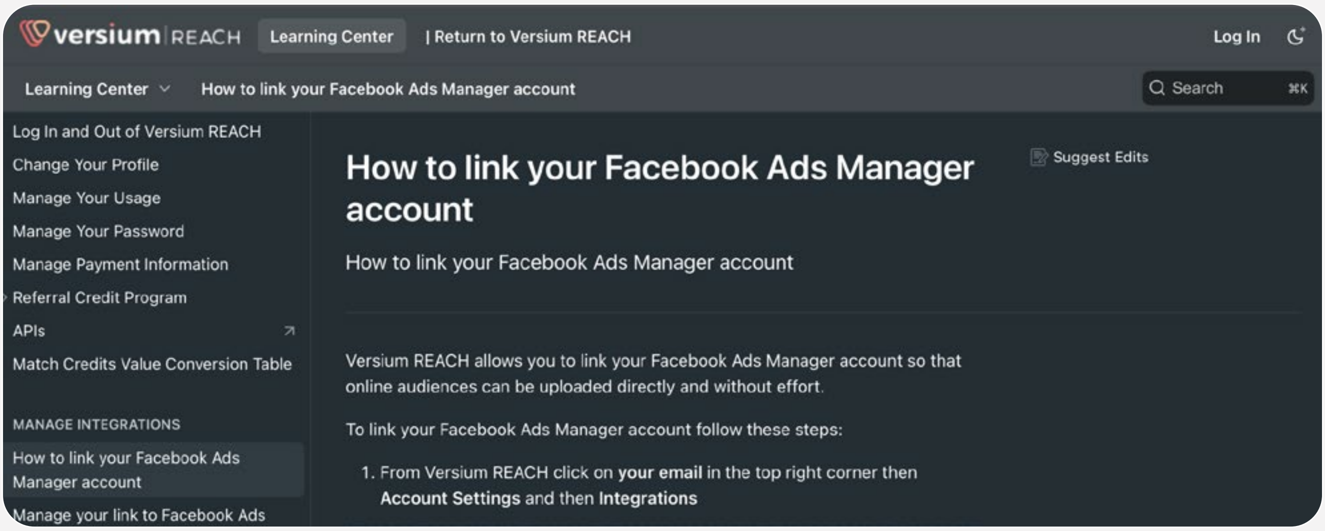
Lifestyle and Interests Attributes

Suggest Edits

Here below is a list of all the Lifestyle and Interests Attributes that Versium REACH may append to your list:

- Magazines
- Reading
- Current Affairs and Politics
- Mail Order Buyer
- Dieting and Weight Loss
- Travel
- Music
- Consumer Electronics
- Arts
- Antiques
- Home Improvement
- Gardening

Figure 26: Examples of Versium's available lists19


versium | REACH
Learning Center
| Return to Versium REACH
Log In
Learning Center
How to link your Facebook Ads Manager account
Search
⌘K
Log In and Out of Versium REACH
Change Your Profile
Manage Your Usage
Manage Your Password
Manage Payment Information
Referral Credit Program
APIs
Match Credits Value Conversion Table
MANAGE INTEGRATIONS
How to link your Facebook Ads Manager account
Manage your link to Facebook Ads
Suggest Edits

How to link your Facebook Ads Manager account

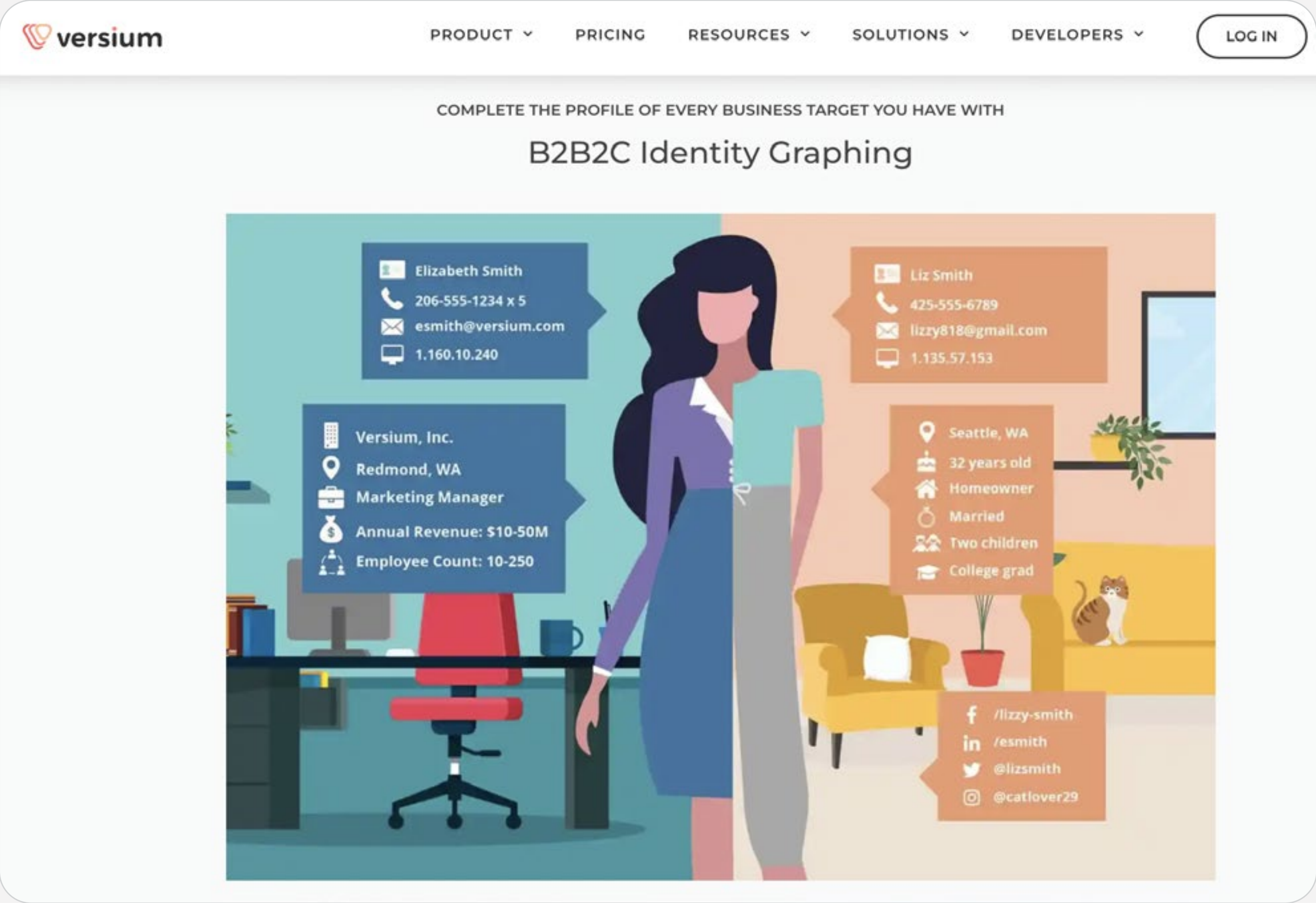
How to link your Facebook Ads Manager account

Versium REACH allows you to link your Facebook Ads Manager account so that online audiences can be uploaded directly and without effort.

To link your Facebook Ads Manager account follow these steps:

1. From Versium REACH click on **your email** in the top right corner then **Account Settings** and then **Integrations**

Figure 27: Excerpts of Versium guidelines on how an advertiser could connect their Meta account to allow them to target Versium's audience[20]



COMPLETE THE PROFILE OF EVERY BUSINESS TARGET YOU HAVE WITH

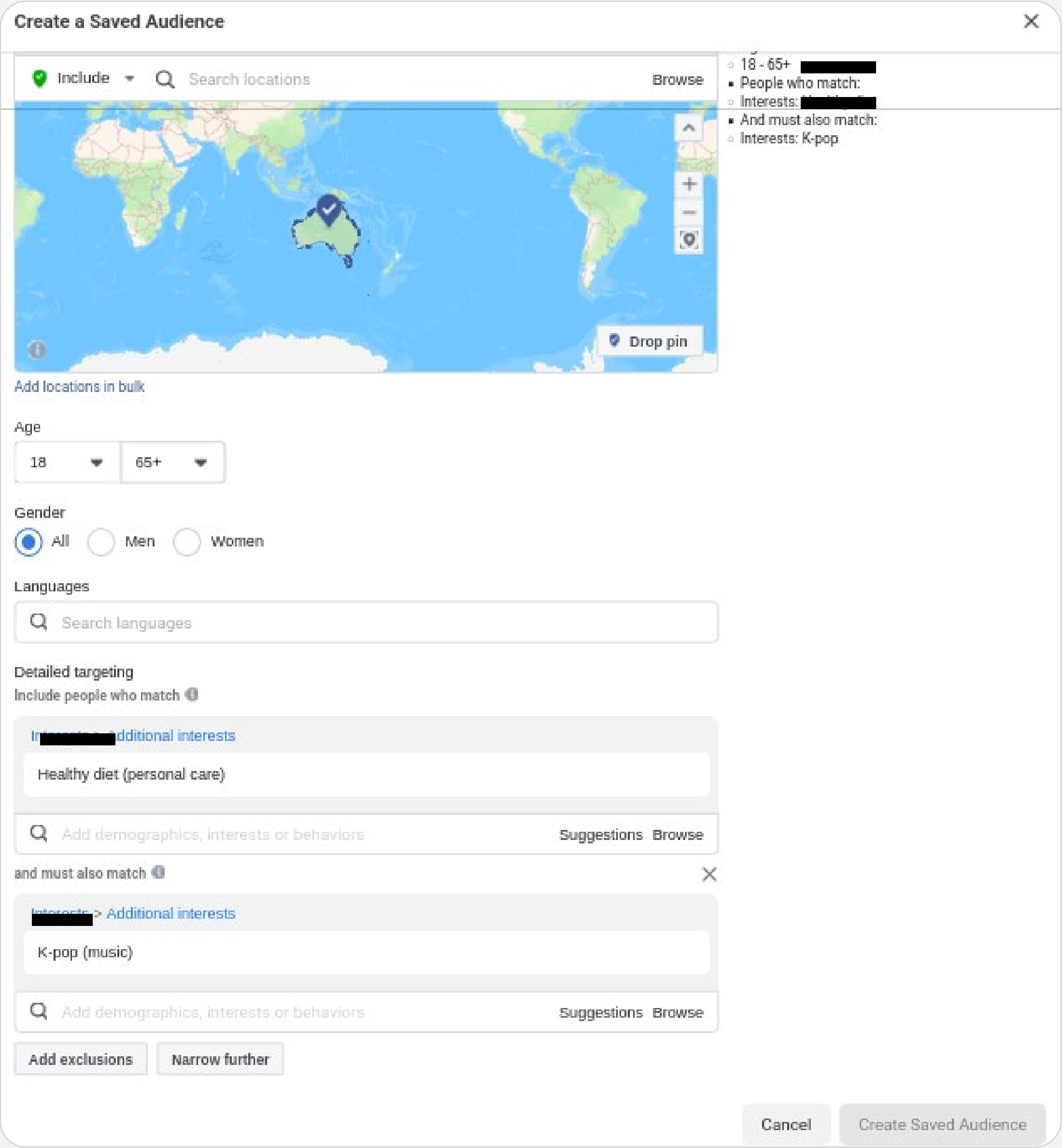
B2B2C Identity Graphing

Figure 28: A diagram of Versium data available about potential customers[21]

- Website traffic: If an advertiser had a relevant website or access to the coding of one—such as a weight loss, mental health or eating disorder website—they would be able to target TikTok users who visited or took specific actions on this website. This could then be expanded to include 'lookalike' end-users.
  - App data: If an advertiser had a relevant app or access to the coding of one—such as a weight loss, mental health or eating disorder website—they would be able to target TikTok users who have downloaded it or taken specific actions on it. This could then be expanded to include 'lookalike' end-users.

» *Using ad interest lists data on Meta*

Meta allows targeting of a broad set of categories using ad interests lists, but these are coarse. There were no lists directly related to pro-eating disorder stars, channels or topics that we uncovered. It might be possible to build target groups by clever combination, e.g. interest in [redacted] and [redacted], and while this is possible (see Figure 28), the coarseness of this probably poses limited risks. We note this is a significant improvement from experiments run by Reset.Tech in 2021, where we found that ads could be targeted to '13-17 years olds' interested in 'extreme weight loss'.[22] Meta announced these reductions in ad interests lists in March 2022.[23]


Create a Saved Audience

- 18 - 65+ [redacted]
- People who match:
  - Interests: [redacted]
- And must also match:
  - Interests: K-pop

Age: 18 – 65+

Gender: ● All ○ Men ○ Women

Languages

Detailed targeting

Include people who match

[redacted] > Additional interests

Healthy diet (personal care)

and must also match

[redacted] > Additional interests

K-pop (music)

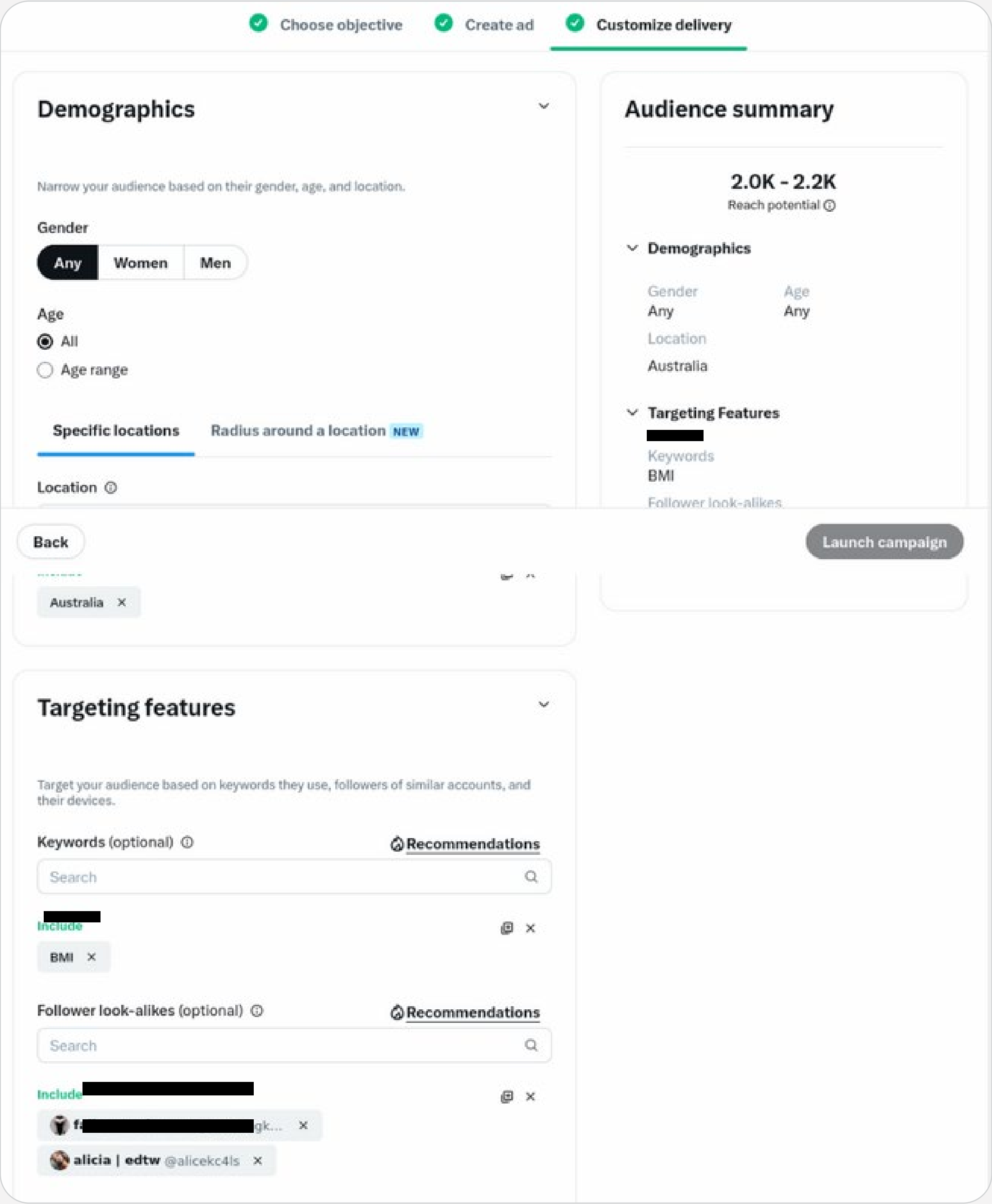
Figure 29: Meta's audience builder using end-user data

X's ad management system

Advertisers could target end-users interested in pro-eating disorder content by using: follower 'lookalikes', and, to a lesser extent key words. These could be comparatively simple and straightforward processes.

» *Using follower lookalikes on X*

One of the easiest ways to target end-users who may be interested in pro-eating disorder content is to create 'follower lookalikes' (i.e. request that the platform target accounts that resemble or share similarities with accounts that follow pro-eating disorder influencers). Follower 'lookalike' categories can be created even from accounts with a low follower count (around 2K), and target groups can be created from them (see Figure 30).



✓ Choose objective ✓ Create ad ✓ Customize delivery

Demographics

Narrow your audience based on their gender, age, and location.

Gender
[● Any] Women Men

Age
● All
○ Age range

Specific locations | Radius around a location NEW

Location

Australia ×

Targeting features

Target your audience based on keywords they use, followers of similar accounts, and their devices.

Keywords (optional) Recommendations
Search

[redacted]
Include
BMI ×

Follower look-alikes (optional) Recommendations
Search

Include [redacted]
f[redacted]gk... ×
alicia | edtw @alicekc4ls ×

Audience summary

2.0K - 2.2K
Reach potential

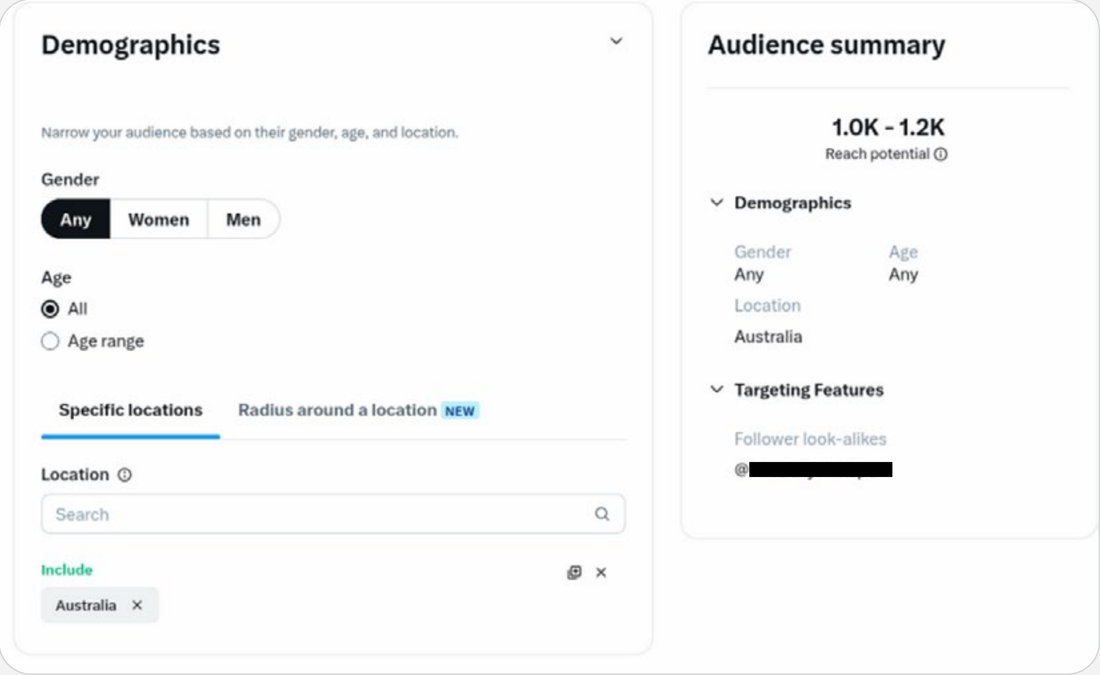
Demographics
Gender: Any
Age: Any
Location: Australia

Targeting Features
[redacted]
Keywords
BMI
Follower look-alikes

Back
Launch campaign

Figure 30: X's follower 'lookalike' functionality, showing how it is possible to target end-users 'like' those who follow pro-eating disorder influencers

These follower 'lookalike' categories can then be refined by selecting geographies (or other demographics, such as age or gender) or adding pro-eating disorder target keywords in the title, such as [redacted] or '[redacted]. For example, from a single account like [redacted], which has 52K followers, it is possible to create a 'lookalike' target group of Australians with over 1K reach (see Figure 31).



Demographics

Narrow your audience based on their gender, age, and location.

Gender

Any | Women | Men

Age

● All

○ Age range

Specific locations | Radius around a location NEW

Location

Search

Include

Australia ×

Audience summary

1.0K - 1.2K

Reach potential

Demographics

Gender: Any

Age: Any

Location: Australia

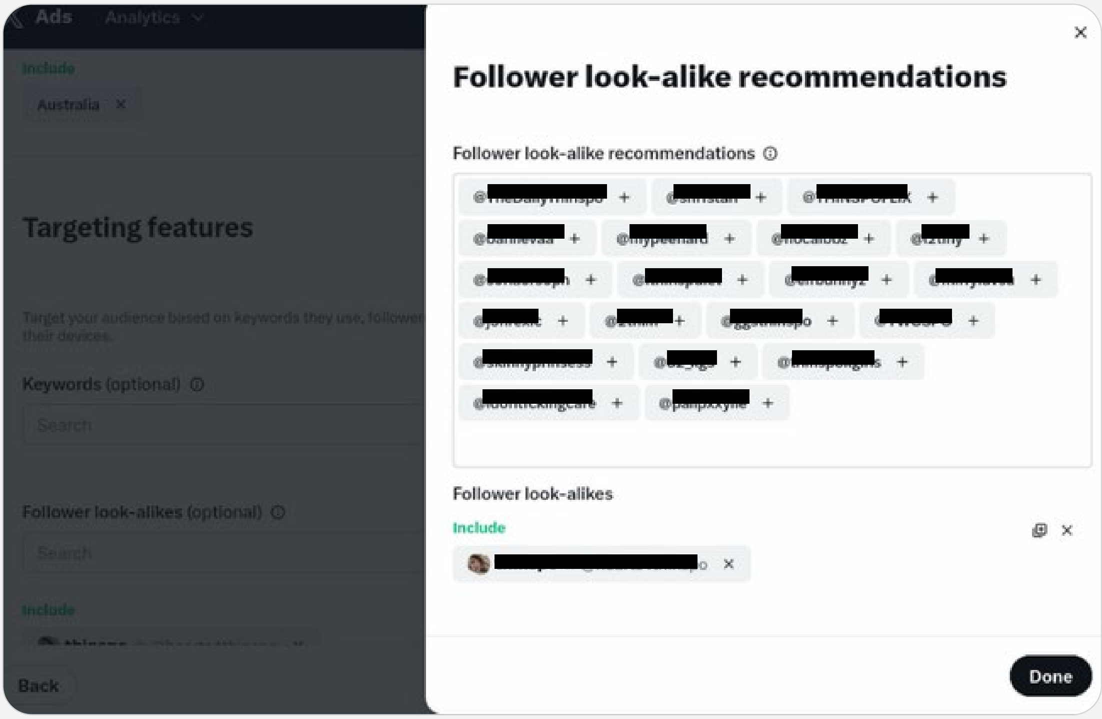
Targeting Features

Follower look-alikes

@[redacted]

Figure 31: X's follower 'lookalike' functionality, showing how it is possible to create even more targeted lists using the 'lookalike' functionality

X even supports advertisers to find pro-eating disorder 'lookalike' end-users. For example, when you suggest creating a 'lookalike' audience for [redacted] and use their recommendations tool, they will suggest similar lookalike accounts, such as [redacted] and [redacted] (see Figure 32).



Include

Australia ×

Targeting features

Target your audience based on keywords they use, follower… their devices.

Keywords (optional)

Search

Follower look-alikes (optional)

Search

Include

Follower look-alike recommendations

Follower look-alike recommendations

[redacted] + [redacted] + [redacted] +
[redacted] + [redacted] + [redacted] + [redacted] +
[redacted] + [redacted] + [redacted] + [redacted] +
[redacted] + [redacted] + [redacted] + [redacted] +
[redacted] + [redacted] + [redacted] +
[redacted] + [redacted] +

Follower look-alikes

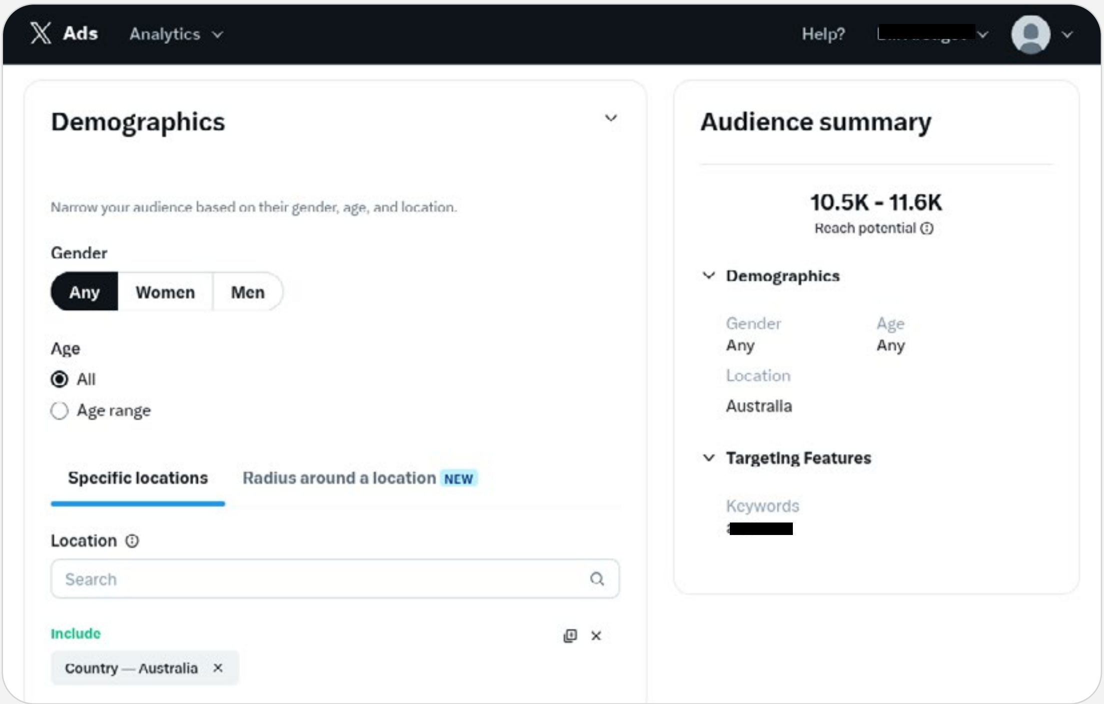
Include

[redacted] ×

Figure 32: X's follower lookalike functionality, showing how it will recommend other pro-eating disorder 'lookalike' end-users to expand the pool

» *Using keywords on X*

X's ad management system blocks advertisers from using some known pro-eating disorder keywords, such as [redacted] and [redacted], to target end-users. However, lesser-known or ambiguous keywords are able to be used to target groups in Australia, including the common keyword[redacted](which may not be blocked because it could also refer to an African news publisher, a Japanese Airline or several NGOs). Using [redacted]as a keyword, X allows advertisers to target 10.50 - 11.6K end-users to target, although these might be Australian end-users with an interest in African news, Japanese flights or NGO news (see Figure 33).



X Ads Analytics ⌄ Help? [redacted] ⌄

Demographics

Narrow your audience based on their gender, age, and location.

Gender

Any | Women | Men

Age

● All

○ Age range

Specific locations | Radius around a location NEW

Location ⓘ

Search

Include

Country — Australia ×

Audience summary

10.5K - 11.6K

Reach potential ⓘ

Demographics

Gender: Any

Age: Any

Location: Australia

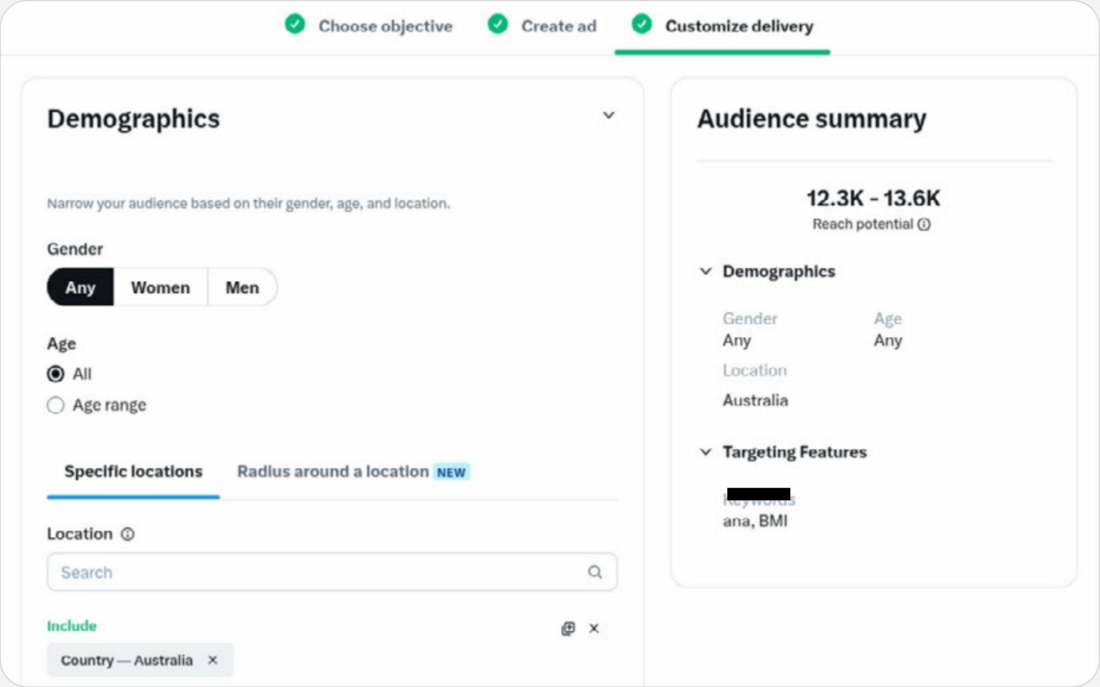
Targeting Features

Keywords

[redacted]

Figure 33: X's keyword functionality, showing how to target 10.5 - 11.6K end-users using the keyword '[redacted]'

Keywords like [redacted] can be combined with other broad keywords like '[redacted]' to target more end-users, but X simply combines the end-users for the keyword [redacted] with end-users for the keyword [redacted]'. 12.3 - 13.6K Australians can be targeted using the keyword combination [redacted] and [redacted] (see Figure 34).



✓ Choose objective ✓ Create ad ✓ Customize delivery

Demographics

Narrow your audience based on their gender, age, and location.

Gender

Any | Women | Men

Age

● All

○ Age range

Specific locations | Radius around a location NEW

Location ⓘ

Search

Include

Country — Australia ×

Audience summary

12.3K - 13.6K

Reach potential ⓘ

Demographics

Gender: Any

Age: Any

Location: Australia

Targeting Features

[redacted]

ana, BMI

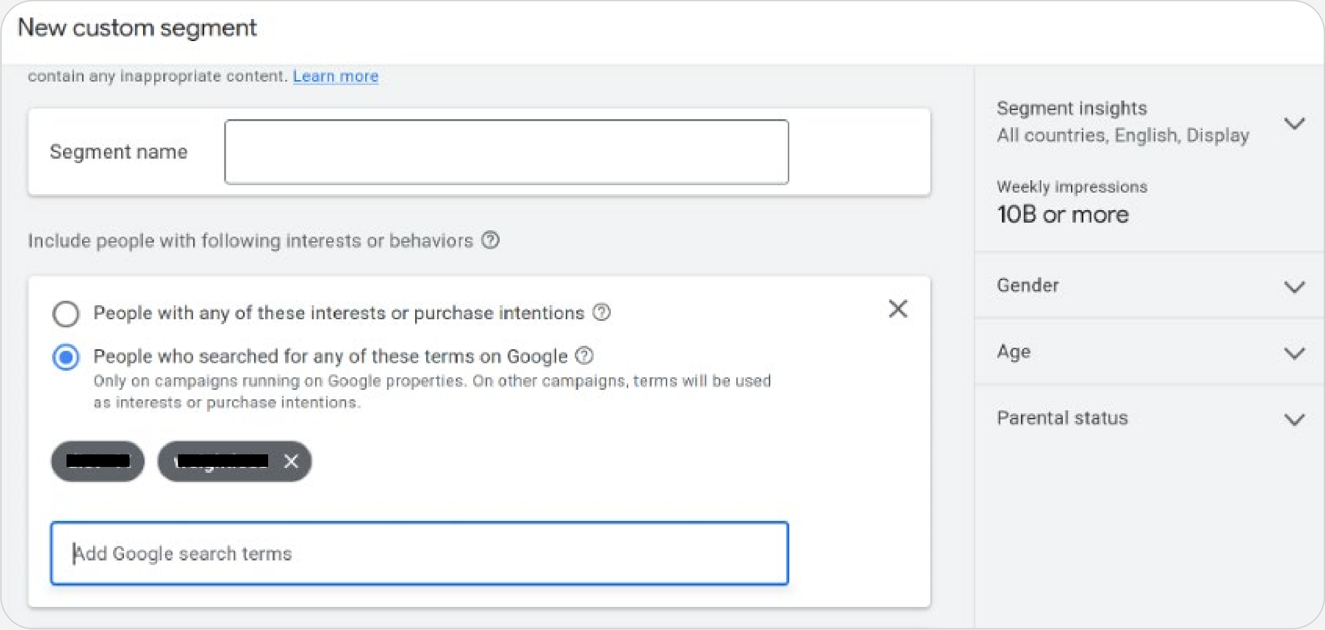
Figure 34: X's keyword functionality, showing how it could target 12.3 - 13.6K end-users using the keywords '[redacted] and[redacted]

Google's ad management system

Advertisers could target end-users interested in pro-eating disorder content by using: Google keyword searches; app download data, and; YouTube channel data. Some of these processes could be comparatively simple or straightforward, and could be very powerful in accuracy.

» *Using Google search terms*

Google allows advertisers to target end-users using keywords, but simple keywords such as [redacted] and [redacted] do not lead to any specific preconfigured target group (see Figure 35).



New custom segment

contain any inappropriate content. Learn more

Segment name

Include people with following interests or behaviors

○ People with any of these interests or purchase intentions

● People who searched for any of these terms on Google
Only on campaigns running on Google properties. On other campaigns, terms will be used as interests or purchase intentions.

[redacted] [redacted]

Add Google search terms

Segment insights
All countries, English, Display

Weekly impressions
10B or more

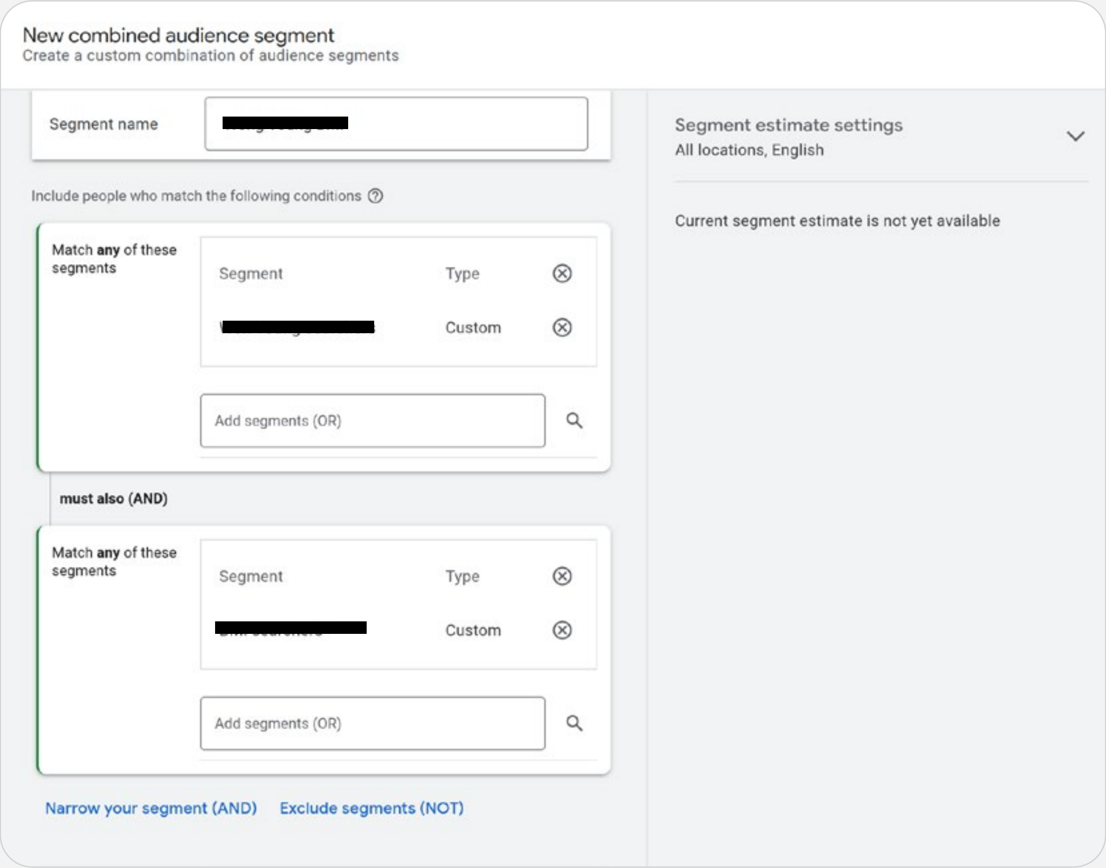
Gender

Age

Parental status

Figure 35: Google's audience builder using keywords

However, advertisers can target end-users using *Google searches* of keywords. These search keywords can be combined to build powerful combinations. For example, it is possible to target end-users who have searched for the [redacted]—who is somewhat of an icon in the eating disorder community—and [redacted]. Combinations like this can be very effective at identifying end-users who are googling pro-eating disorder materials online (see Figure 36).



New combined audience segment
Create a custom combination of audience segments

Segment name: [redacted]

Include people who match the following conditions

Match **any** of these segments

Segment	Type
[redacted]	Custom

Add segments (OR)

must also (AND)

Match **any** of these segments

Segment	Type
[redacted]	Custom

Add segments (OR)

Narrow your segment (AND) Exclude segments (NOT)

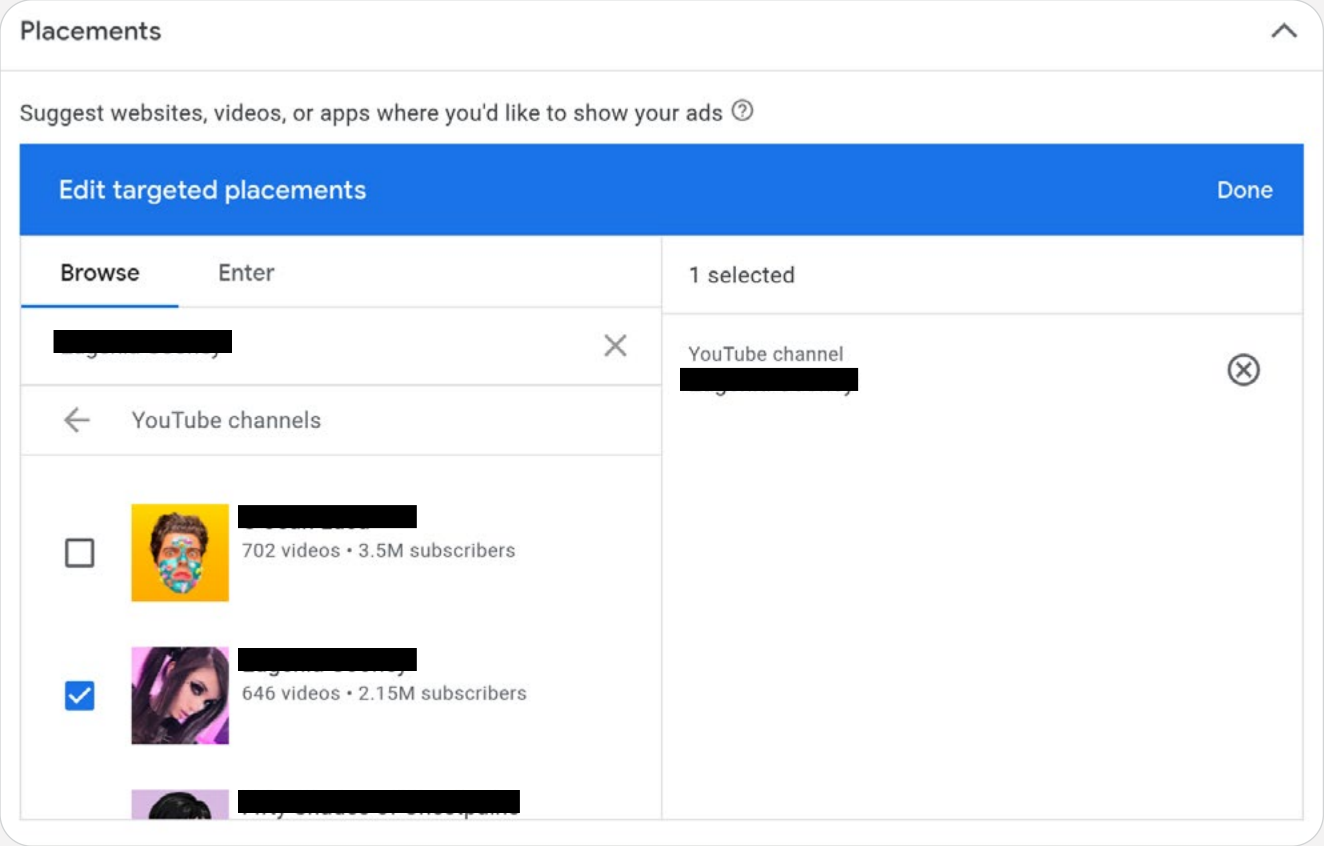
Segment estimate settings
All locations, English

Current segment estimate is not yet available

Figure 36: Google's audience builder showing how to target end-users who have made specific keyword searches

» *Using YouTube channel data*

Advertisers can also target end-users interested in pro-eating disorder content using YouTube data. Advertisers can identify pro-eating disorder YouTube channels, or even eating disorder support channels, and target ads within these channels. There are some well known pro-eating disorder YouTube channels, such as [redacted] which has 2.2M followers (see Figure 37). While any ads placed in these channels will be subject to ad review processes, our ad approval experiment shows that this system is far from effective.



Placements

Suggest websites, videos, or apps where you'd like to show your ads

Edit targeted placements — Done

Browse | Enter

[redacted]

YouTube channels

□ [redacted] 702 videos • 3.5M subscribers

☑ [redacted] 646 videos • 2.15M subscribers

[redacted]

1 selected

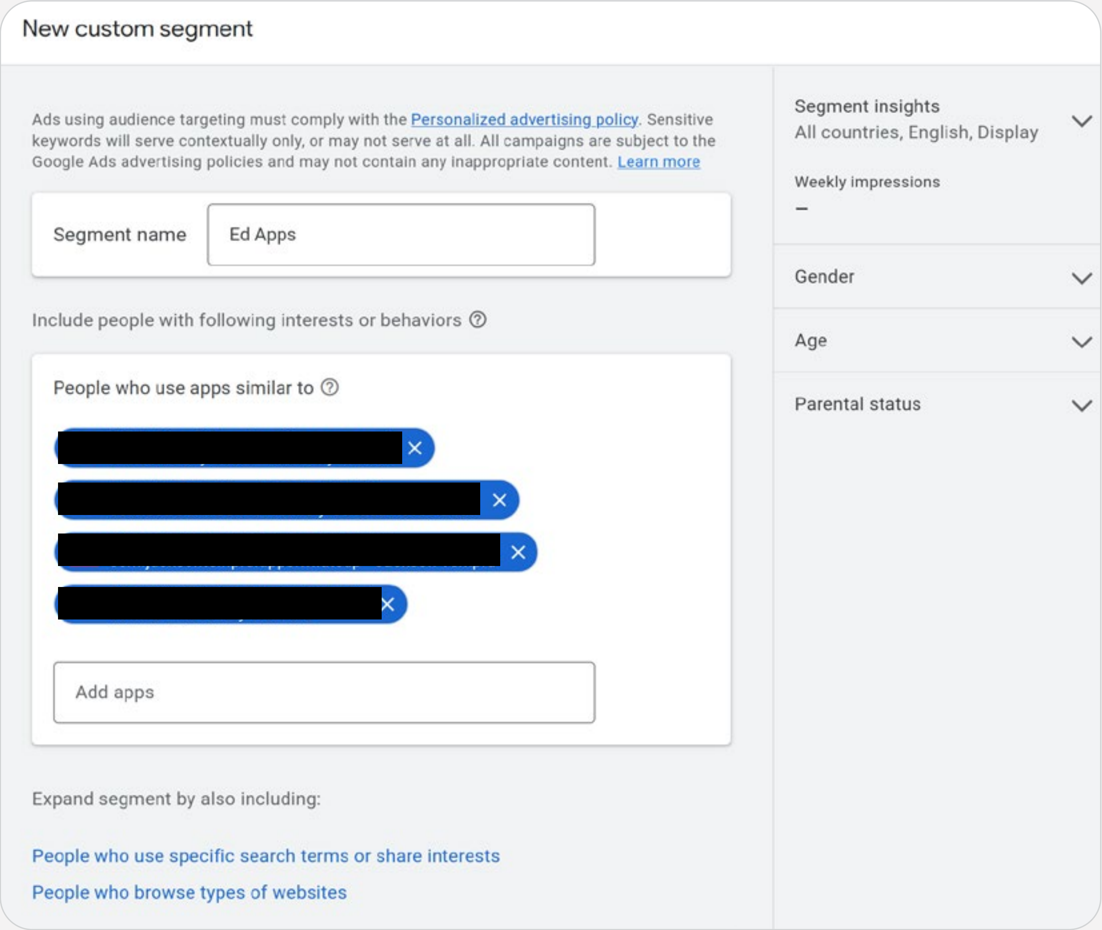
YouTube channel [redacted]

Figure 37: Google's audience builder showing how ads can be placed in a pro-eating disorder YouTube channel

» *Using app download data*

Google Ads has data about app downloads and allows advertisers to target end-users based on which apps they have downloaded from the Google Play store. Advertisers can probably target those who have downloaded eating disorders apps or other mental health apps (see Figure 38). Figure 39 highlights why we are saying 'probably'; when we constructed a segment of end-users who downloaded four common mental health and eating disorder apps, we did not get an estimate of how many users this would reach. This is most likely because our research account on Google was new so Google would not provide this information, but it means we are unable to confirm if there is a block on these apps. Our account was also not given estimates about how many users could be reached by targeting users who had downloaded mainstream apps such as Instagram, which suggests that this is not a block on eating disorder apps specifically.

To be clear, the very apps that are often created to support people affected by eating disorders could just as easily be used to render them vulnerable to advertisers. It is worth noting that the mental health app market is largely unregulated, and not all mental health apps will have been created to offer effective support.



New custom segment

Ads using audience targeting must comply with the Personalized advertising policy. Sensitive keywords will serve contextually only, or may not serve at all. All campaigns are subject to the Google Ads advertising policies and may not contain any inappropriate content. Learn more

Segment name: Ed Apps

Include people with following interests or behaviors ⓘ

People who use apps similar to ⓘ

[redacted] ✕

[redacted] ✕

[redacted] ✕

[redacted] ✕

Add apps

Expand segment by also including:

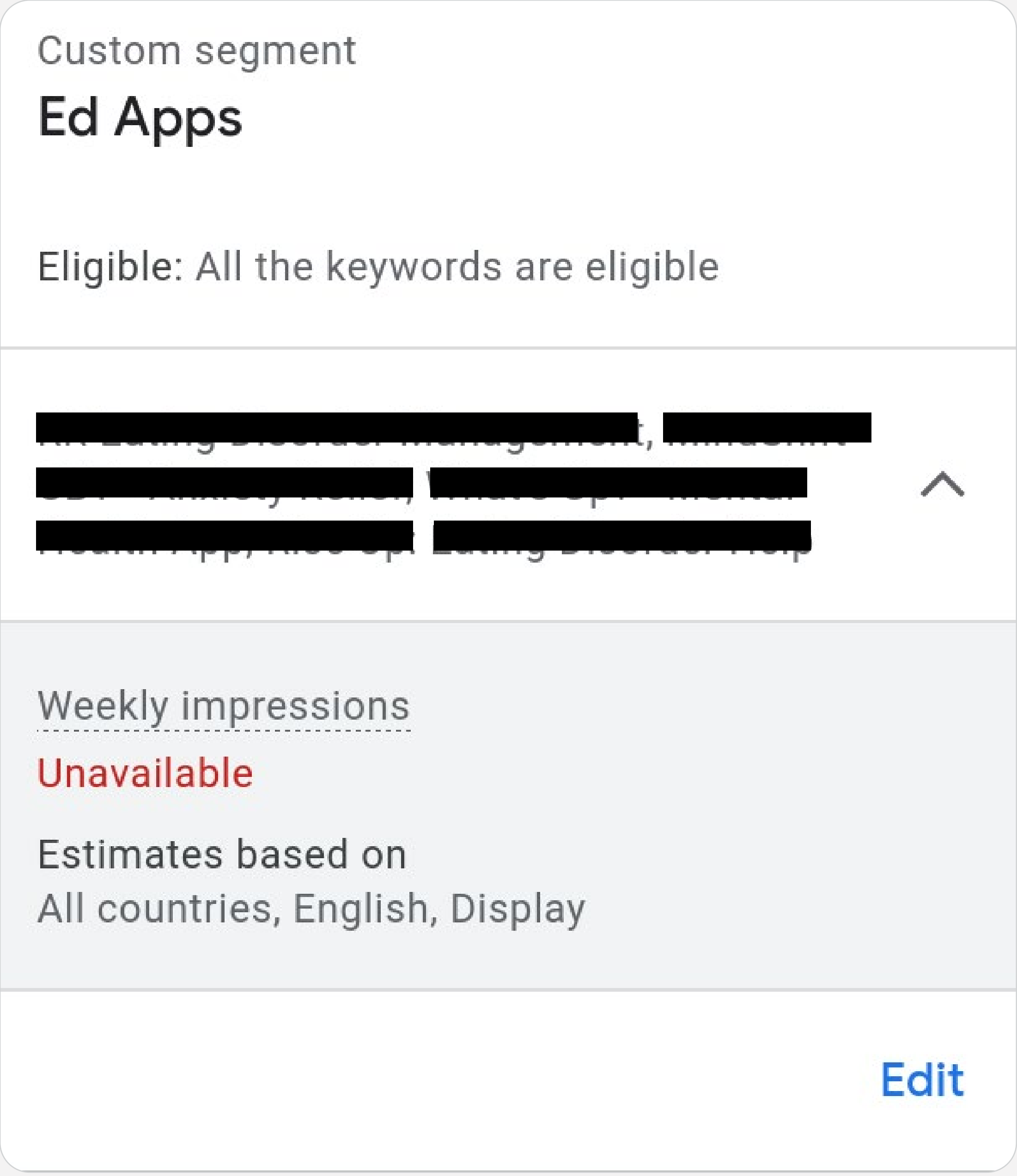
People who use specific search terms or share interests

People who browse types of websites

Segment insights
All countries, English, Display

Weekly impressions
–

Figure 38: Google's audience builder showing how end-users appear to be able to be targeted by the apps they have downloaded



Custom segment

Ed Apps

Eligible: All the keywords are eligible

[redacted]

Weekly impressions

Unavailable

Estimates based on

All countries, English, Display

Figure 39: Google's audience builder showing how end-users who have downloaded the four mental health and eating disorder apps in figure 38 can be targeted, but without showing the reach of this approach

Appendices

ENDNOTES

1 TikTok 2024 *Community Guidelines* https://www.tiktok.com/community-guidelines/en/.

2 TikTok 2024 *Mental and Behavioural Health Guidelines* https://www.tiktok.com/community-guidelines/en/mental-behavioral-health/

3 Instagram 2024 Community Guidelines https://help.instagram.com/477434105621119/?helpref=hc_fnav.

4 Meta 2024 *Suicide and Self Injury* https://transparency.fb.com/en-gb/policies/community-standards/suicide-self-injury/?source=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards%2Fsuicide_self_injury_violence.

5 X 2024 *Suicide and Self Harm policy* https://help.twitter.com/en/rules-and-policies/glorifying-self-harm.

6 See InsideOut 2024 *InsideOut Institute for Eating Disorders* https://insideoutinstitute.org.au/

7 TikTok 2024 *Advertising Policies: Industry Entry - Oceania* https://ads.tiktok.com/help/article/tiktok-advertising-policies-industry-entry-oceania?lang=en

8 Meta 2024 *Advertising Standards* https://transparency.fb.com/en-gb/policies/ad-standards/objectionable-content/personal-health-and-appearance/

9 Meta 2024 *Advertising Standards* https://transparency.fb.com/en-gb/policies/ad-standards/objectionable-content/personal-health-and-appearance/

10 X 2024 *Ads Policies* https://business.twitter.com/en/help/ads-policies/about-twitter-ads-approval.html

11 X 2024 *Ads Content Policies: Inappropriate Content* https://business.twitter.com/en/help/ads-policies/ads-content-policies/inappropriate-content.html

12 Google 2024 *Google Advertising Policies: Inappropriate Content* https://support.google.com/adspolicy/answer/6015406?hl=en

13 Data would need to include a static mobile advertising ID, which is the norm.

14 For more information about data for sale uncovered in the Xandr file, see Reseet.Tech 2023 *Australians for Sale* https://au.reset.tech/news/coming-soon-australians-for-sale-report/

15 Meta 2024 *Data feed fields and specifications for catalogues in Commerce Manager* https://www.facebook.com/business/help/120325381656392?id=725943027795860

16 Data would need to include a static mobile advertising ID, which is the norm.

17 For more information about data for sale uncovered in the Xandr file, see Reset.Tech 2023 *Australians for Sale* https://au.reset.tech/news/coming-soon-australians-for-sale-report/

18 More details about Versium can be found at Versium 2024 Versium https://versium.com/ and Versium 2024 Getting started https://reach-help.versium.com/

19 Versium nd *Lifestyle and Interests attribute* https://reach-help.versium.com/docs/lifestyle-and-interests-attributes

20 Versium nd *How to link your Facebook ads manager account* https://reach-help.versium.com/docs/how-to-link-your-facebook-ads-manager-account

21 Versium nd *Why Versium* https://versium.com/why-versium

22 Reset.Tech 2021 *Profiling Children for Advertising* https://au.reset.tech/news/profiling-children-for-advertising-facebooks-monetisation-of-young-peoples-personal-data/

23 Meta 2022 *Removing Certain Ad Targeting Options and Expanding Our Ad Controls* https://www.facebook.com/business/news/removing-certain-ad-targeting-options-and-expanding-our-ad-controls/

Download the digital report and appendices at https://au.reset.tech/news/

‘Not Just Algorithms: Assuring User Safety Online with Systemic Regulatory Frameworks’

Reset.Tech Australia,
with support from Dr Hannah Jarman

March 2024

POLICY REPORT

Reset.
AUSTRALIA

Version 1.0 240323