

Reset.

An Evaluation of X's Processes for Risks to Minors

I

Summary

Summary

This report documents an evaluation of systems on X (Twitter) to assess the risks posed to minors. These systems include:

- X's content moderation system;
- The platform's understandability for younger users;
- X's algorithmic content recommender system;
- X's ad manager systems.

We identified several issues that may not comply with the Digital Services Act, such as:

- X under-moderates pro-restrictive eating disorder content, pro-suicide, and/or pro-self-harm materials;
- There is a muted response to these materials when X becomes aware of them via the user-reporting system, and X fails to respond to the majority of pro-restrictive eating disorder content and pro-suicide and/or pro-self-harm materials when they become aware of it;
- X's content recommender system may promote pro-suicide and/or self-harm content and pro-eating disorder content to young people;
- X's content recommender system disproportionately promotes hashtags associated with pro-suicide and/or self-harm content;
- A 13-year-old may not understand the design and functioning of X at the point of signing up because the Cookie policy they are directed to read may not be available in their first language, and dark patterns are deployed in the sign-up process.

Table of contents

Introduction	5
An Evaluation of X's Content Moderation Systems in Creating and Perpetuating Risks to Minors	8
Methodology	9
Findings	10
X's response to pro-suicide and/or self-harm material	10
X's response to pro-restrictive eating disorder material	11
Limitations	11
Conclusion	11
An Evaluation of X's Algorithmic Recommender Systems and Violative Content	12
Methodology	13
Findings	15
Harmful content recommended to sock puppet accounts	15
Promotion of hashtags associated with harmful content	16
Conclusion	18
An Evaluation of Understandability of X for Young Users, Including Dark Patterns	19
Methodology	20
Findings	21
A typology of dark patterns in the sign-on experience	22
Dark patterns discovered in the sign-on experience	22
Accessibility and comprehensibility of policies	24
Conclusion	26
Appendices	27
Appendix 1: X's content moderation guidelines	28
X's community guidelines on suicide and/or self-harm	28
X's community guidelines on eating disorder content	30
Appendix 2: Examples of X's monitored content	33
Pro-suicide and/or self-harm	34
Pro-eating disorder	35
Appendix 3: Examples of X's monitored content	36
Appendix 4: X's sign-on process	40

III

Introduction

Introduction

This report documents an evaluation of systems on X (Twitter) to assess the risks posed to minors. These systems include:

- X's content moderation system;
- The platform's understandability for younger users;
- X's algorithmic content recommender system;
- X's ad manager systems.

The Digital Service Act (DSA) aims to provide additional protections for children and young people under 18 years old in the digital sphere.

- Recital 71 states that “the protection of minors is an important policy objective of the Union.” Platforms are considered accessible to minors when:
 - Its terms and conditions permit minors to use the service;
 - Its service is directed at or predominantly used by minors;
 - Where the provider is otherwise aware that some of the recipients of its service are minors, for example, because it already processes personal data of the recipients of its service revealing their age for other purposes.
- Recital 71 goes on to state, “Providers of online platforms used by minors should take appropriate and proportionate measures to protect minors, for example, by designing their online interfaces or parts thereof with the highest level of privacy, safety and security for minors by default where appropriate or adopting standards for protection of minors, or participating in codes of conduct for protecting minors. They should consider best practices and available guidance, such as that provided by the communication of the Commission on A Digital Decade for children and youth: the new European strategy for a Better Internet for Kids (BIK+). Providers of online platforms should not present advertisements based on profiling using personal data of the recipient of the service when they are aware with reasonable certainty that the recipient of the service is a minor.”
- Recital 81 further indicates that very large online platforms should consider, for example, “how easy it is for minors to understand the design and functioning of the service, as well as how minors can be exposed through their service to content that may impair minors' health, physical, mental, and moral development.” Such risks may arise, for example, in relation to the design of online interfaces that intentionally or unintentionally exploit the weaknesses and inexperience of minors or which may cause addictive behavior.
- Recital 84 explains that in assessing systemic risk—which includes risks to minors—“providers of very large online platforms and of very large online search engines should focus on the systems or other elements that may contribute to the risks, including all the algorithmic systems that may be relevant, in particular their recommender systems and advertising systems, paying attention to the related data collection and use practices.”
- In addition, Article 34 places additional requirements on Very Large Online Platforms (VLOPs) and Very Large Online Search Engines to assess the risks their services pose to children's rights. Specifically, Article 34(1)(d) DSA requires VLOPs to undertake risk assessments, including “any actual or foreseeable negative effects in relation to [...] minors.” Article 34(2)(b) DSA explicitly states that algorithmic recommender systems, content moderation systems, enforcement of terms and conditions, and advertising systems be considered.

● An Evaluation of X's Processes for Risks to Minors

This report examines X's adherence to the requirements outlined in the mentioned recitals and articles. Specifically, it assesses four systems on X for compliance:

- 1 **Content moderation systems:** A method is proposed for testing and evaluating these with regards to creating risks to minors. Specifically, it describes the method used to evaluate if platforms remove content that is harmful to minors when they become aware of it through user-reports. It describes the methods and presents findings from a September 2023 experiment around reporting and monitoring two bodies of content that were assessed by a clinical psychologist and deemed to be harmful to children:
 - a. Pro-suicide and/or self-harm content;
 - b. Pro-restrictive eating disorder content.
- 2 **“Understandability” for young people:** A simple method is developed to evaluate “understandability” for young people. It assesses for dark patterns, meaning platforms' design decisions that cumulatively nudge users to accept default choices that may be against their interests. It describes the methods and presents findings from a September 2023 analysis of three platforms, based on an analysis of the user-journey when new accounts for minors are created.
- 3 **Algorithmic content recommender systems:** An assessment is performed to determine whether content is recommended that violates X's community guidelines. Specifically, this evaluation explores if the recommender system will promote pro-suicide and/or self-harm content, as well as pro-eating disorder content to teenage sock puppet accounts, examining if it overly promotes hashtags associated with pro-suicide and/or pro-self-harm content. We describe a process for evaluating this.
- 4 **Ad manager system:** A method is employed to test if the platform allows advertising to minors based on profiling.

III

*An Evaluation of X's Content
Moderation Systems in Creating
and Perpetuating Risks to Minors*

An Evaluation of X's Content Moderation Systems in Creating and Perpetuating Risks to Minors

Research questions:

- 1 Does X adequately moderate pro-suicide and/or self-harm material when they become aware of it?
- 2 Does X adequately moderate pro-restrictive eating disorder material when they become aware of it?

Methodology

The research involved five steps:

1 Developing criteria to define harmful material.

- This research explored two bodies of content posing psychological and physiological risks to minors: pro-suicide and/or self-harm material, and pro-restrictive eating disorder material.
- We used the community guidelines for each platform to develop a coding schema to classify content (see Appendix 1 for more details). This ensures that only content violating X's Terms of Service was included in this research. Each piece of content, according to their guidelines, should warrant a content-moderation action from X.

2 Identifying pro-suicide and/or self-harm material.

Using simple searches, we identified content on X that met our criteria and had not been labelled by the platform already. We consulted a clinical psychologist who assessed each piece of identified content, confirming its risk to young consumers. Material not deemed harmful by the psychologist was excluded.

Total identified content:

- Pro-suicide and self-harm content: 96 pieces
- Pro-restrictive eating disorder content: 111 pieces

See Appendix 2 for examples of these bodies of content.

3 Monitoring content pre-reporting.

We tracked this content for two weeks noting:

- View counts and growth rates;
- Labelling or warning rates to observe if X labelled any content during these two weeks. Considered labelled if an age-restriction warning, sensitivity filter, or any other flag was placed on it;
- Take-down rates to check if X removed any content during these two weeks.

4 Reporting the content.

We reported each piece of content as suicide and self-harm, or restrictive eating disorder content violating the Terms of Service to the platform.

5 Monitoring content post-reporting.

After reporting, we tracked this content for two further weeks noting:

- View counts and growth rates;
- Labelling or warning rates to observe if any content was labelled by the platforms during these two weeks. Considered labelled if an age-restriction warning, sensitivity filter, or any other flag was placed on it;
- Take-down rates to check if any content was removed by the platforms during these two weeks.

According to our analysis of the platform's community guidelines (see Appendix 1), X should delete pro-suicide and/or self-harm content and pro-eating disorder content when they become aware of it. In practice, we often see platforms label, add sensitivity filters, or age filters to this body of materials. We, therefore, also assess these.

Below, we describe what we found over four weeks of monitoring.

Findings

X's response to pro-suicide and/or self-harm material

X does not adequately label or demote pro-suicide and/or pro-self-harm content.

Removal appears to be the most common response to pro-suicide and/or pro-self-harm material, but X's reactions to reporting are inadequate. Most of the content remained available and unlabelled, even after user-reporting.

Over two weeks monitoring	X
Pre reporting removal rate. This is the % of content that was removed during the two weeks before we reported it. It 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 rates.	6.25%
Post reporting removal rate. This is the % of content that was removed within 2 weeks after we reported it.	13.33%
Effect of reporting on removal rate	+7.08%
Pre reporting labelling or warning rate. This is the % of content that was labelled during the two weeks before we reported it. It may have been reported by other users, but represents the best estimate of organic labelling rates.	0%
Post reporting labelling or warning rate. This is the % of content that was labelled within 2 weeks after we reported it.	0%
Effect of reporting on labelling rate	No change
Pre reporting growth rate. This is the average growth rate of content over two weeks before we reported it (week-on-week).	2.77% growth week-on week
Post reporting growth rate. This is the average growth rate of content over two weeks after we reported it (week-on-week).	2% growth week-on week
Effect of reporting on growth rate	-0.77%

● An Evaluation of X's Processes for Risks to Minors

X's response to pro-restrictive eating disorder material

X does not appear to adequately label or demote pro-restrictive eating disorder content.

Removal appears to be the most common response to pro-eating disorder material, but X's reactions to reporting are inadequate. Most of the content remained available and unlabelled, even after user-reporting.

Over two weeks monitoring	X
Pre reporting removal rate. This is the % of content that was removed during the two weeks before we reported it. It 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 rates.	2.7%
Post reporting removal rate. This is the % of content that was removed within 2 weeks after we reported it.	6.48%
Effect of reporting on removal rate	+3.78%
Pre reporting labelling or warning rate. This is the % of content that was labelled during the two weeks before we reported it. It may have been reported by other users, but represents the best estimate of organic labelling rates.	0%
Post reporting labelling or warning rate. This is the % of content that was labelled within 2 weeks after we reported it.	0%
Effect of reporting on labelling rate	No change
Pre reporting growth rate. This is the average growth rate of content over two weeks before we reported it (week-on-week).	2.57%
Post reporting growth rate. This is the average growth rate of content over two weeks after we reported it (week-on-week).	3.55%
Effect of reporting on growth rate	+0.99%

Limitations

When content is removed, the reasons for why it was removed can be unclear. It could have been removed by the users, the user may have deleted their account or switched to private, or the content or the account may have been removed by the platform itself.

The estimations for removal rates, therefore, represent the highest-end estimations of removal rates by platforms.

Conclusion

- X under-moderates both pro-restrictive eating disorder content and pro-suicide and/or pro-self-harm materials.
- There is a muted response to these materials when X becomes aware of them via user-reporting systems. X does not appear to adequately remove, label, or demote pro-restrictive eating disorder content, nor pro-suicide and/or pro-self-harm materials.

IV

An Evaluation of X's Algorithmic Recommender Systems and Violative Content

An Evaluation of X's Content Moderation Systems in Creating and Perpetuating Risks to Minors

Research questions:

- 1 Does X's content recommender system promote violative content to minors' feeds?

Methodology

This research involved two processes and six steps:

Process 1: Evaluation using sock puppet methods

- 1 We created 6 sock puppet accounts for young people on X:
 - a. Slovenia: 16 years old
 - b. Italy: 14 and 16 years old
 - c. Finland: 13 and 16 years old
 - d. Germany: 16 years old

For Italy, the minimum age required for opening an X account is 14. For Slovenia and Germany, no 13-year-old sock puppet accounts were created because the two countries require users to be at least 16 to have an X account.

- 2 We primed each sock puppet account. The sock puppet accounts were primed with violative content collected for 20 minutes each. Thirteen- or 14-year-old accounts were primed with self-harm and pro-suicide content, while 16-year-olds were primed with eating disorder content. The reason for priming is to evaluate whether the platform's algorithms demote or keep recommending similar kinds of violative content in the feed despite user engagement with such content.
- 3 We then monitored the X main feed for each account. Two days' feed in each sock puppet account between 15 September and 21 September were evaluated to identify the presence of violative content similar to the ones during priming. For each account, approximately 500 segments of posts in the feed were examined.

Process 2: Evaluation using domain demotion methods

- 1 We used a domain demotion methodology¹ to detect algorithmically downranked content on social media platforms. We used a source list of violative content collected for the "Evaluation of X's Content Moderation Systems in Creating and Perpetuating Risks to Minors" evaluation, as described above, which was all confirmed as harmful by a clinical psychologist. We evaluated the posts of all the accounts who shared violative content to create a baseline consisting of posts from January to September 2023. The dataset includes metrics such as likes, retweets, comments, bookmarks (collectively "engagement") and views (exposure). For each user, we calculated a baseline for engagement and exposure metrics. We calculated Z-scores to standardize the performance metrics of individual posts against the user-specific baseline. The Z-score captures how far a given data point is from the mean, measured in standard deviations. This standardization allows us to identify outliers and facilitates the comparison of data sets across different users.

1 <https://drive.google.com/file/d/19eUj8bUakT3b5SPto9jbLQp8POMttZFD/view>

● An Evaluation of X's Processes for Risks to Minors

- 2 We applied a log transformation to normalize the data to address the skewness and the power law common on social media, according to which some posts go viral, but the majority receive few engagements. Posts with Z-scores deviating significantly from zero, especially those confirmed through statistical tests, are considered candidates for algorithmic downranking. We used statistical tests to investigate further the significance of observed downranking trends. The Welch Two Sample t-test was used for this purpose. A low p-value (below 0.05) was considered strong evidence of a systematic difference between samples.
- 3 To assess whether X implemented measures to demote harmful content, we used ChatGPT to identify and classify² self-harm-related hashtags, followed by human review. We validated the category selection and compared engagement for self-harm-related hashtags to a baseline of all other posts.

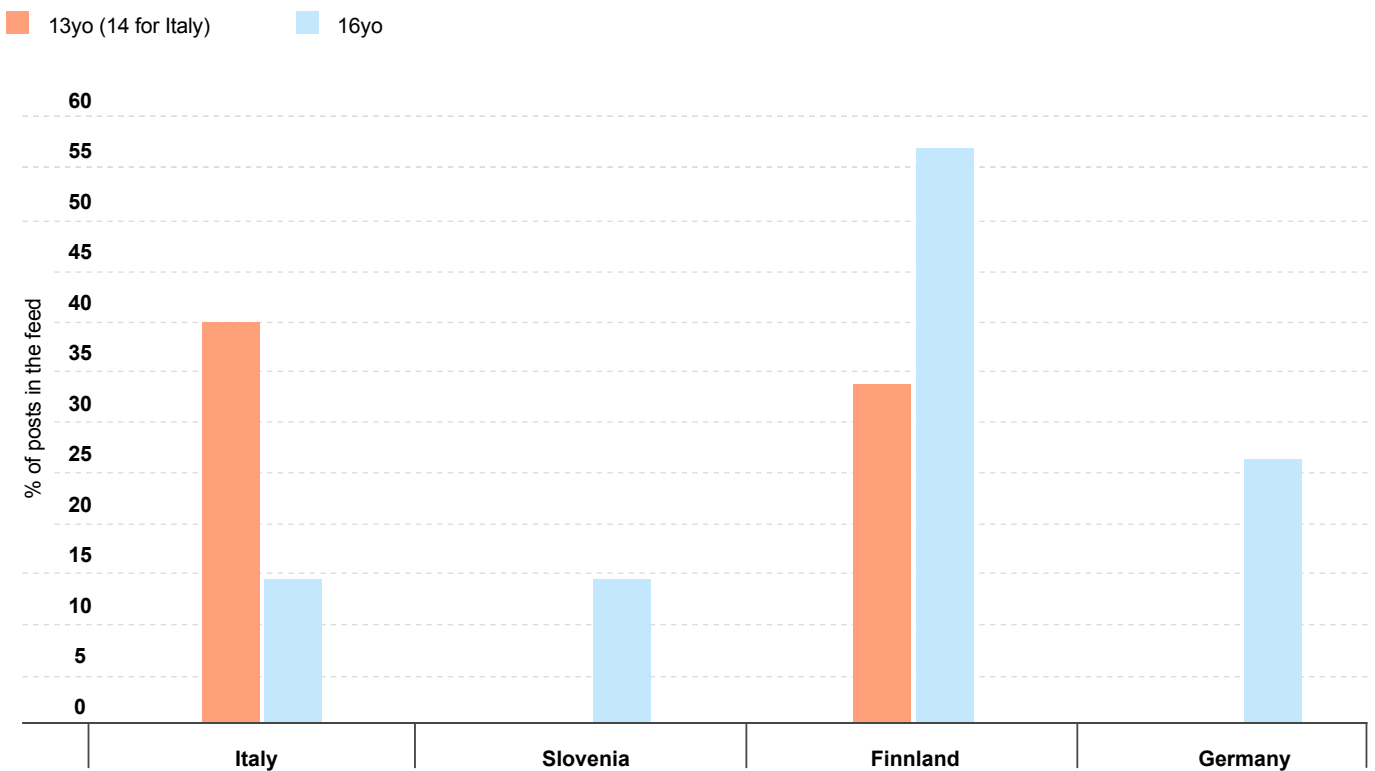
2 For reference, the ChatGPT interaction can be found here: <https://chat.openai.com/share/704f4ded-f923-439a-aa41-2dd0dd86aad2>

Findings

Harmful content recommended to sock puppet accounts

We found that the feeds of teen accounts on X (Twitter) contained explicit self-harm or eating disorder content. For example, 35 posts (8.1%) from a 14-year-old Italian's X account display self-harm; 57 posts (13.2%) from a 16-year-old Finn's X account feature eating disorders. In this experiment, X did not restrict showing self-harm or eating disorder content in teen users' feeds. Instead, it reinforced teen users' proclivity for violative content through its algorithms and consistently recommended such content in the feed.

Most of the feed content in the X sock puppet accounts comes from auxiliary accounts of those the sock puppet accounts follow, as the X signup process requires the user to follow at least one account. For example, the Italy 14-year-old X account followed BTS (@BTS_Official), and therefore, posts from fan accounts and promotional accounts such as Sel7 (@BTStran- slation_) and BTS Charts & Awards (@btschartstudio) were prominent among feeds in the posts. However, besides recom- mending based on the interests selected and the account followed, X's recommender system pushed harmful content into the sock puppet accounts' feeds.



Graph 1: number of posts in the evaluated feed of the X (Twitter) sock puppet accounts

See appendix three for examples of content that was served to sock puppet accounts.

● An Evaluation of X's Processes for Risks to Minors

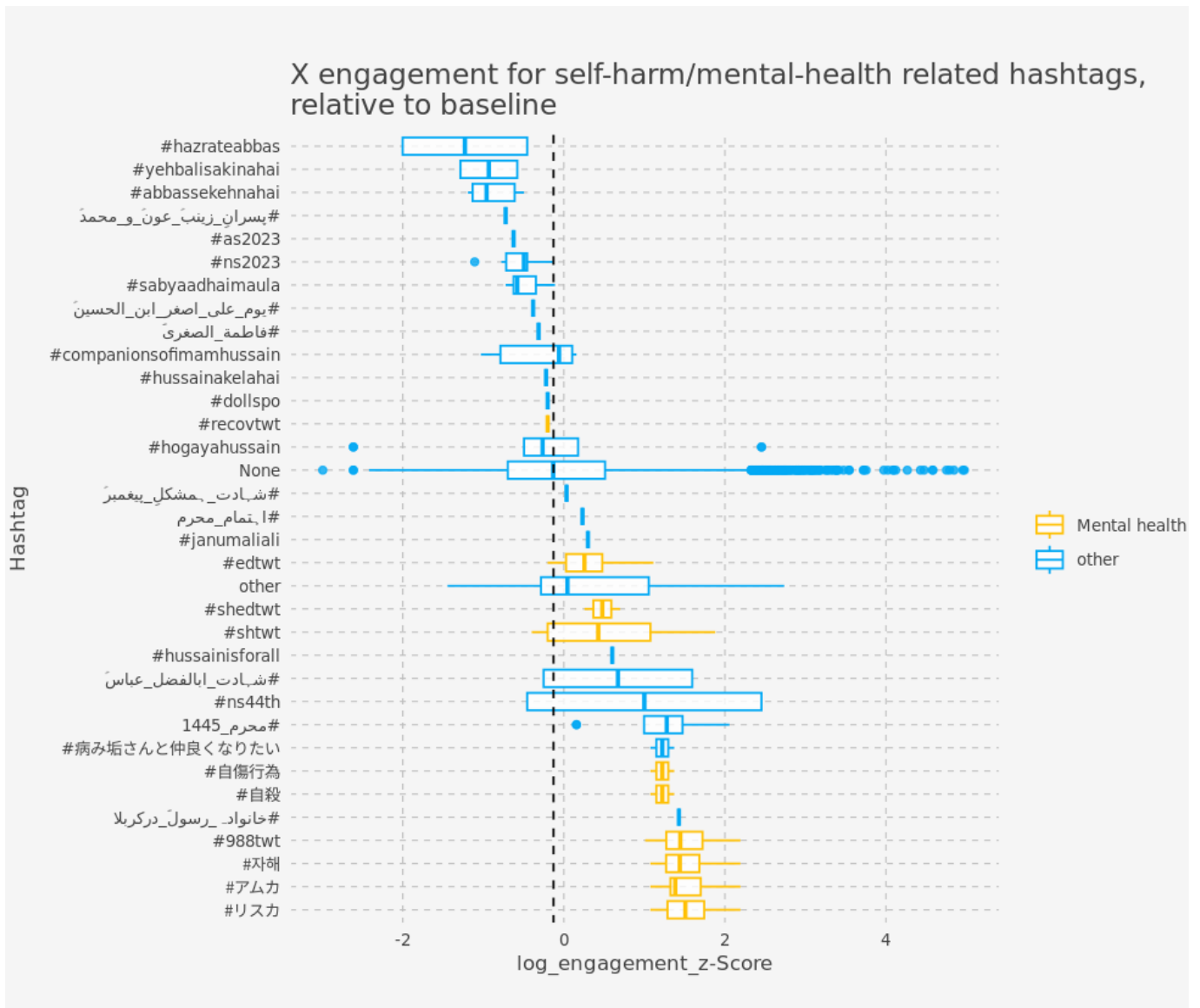
Promotion of hashtags associated with harmful content

We identified the following hashtags as related to self-harm, mental health, and recovery communities:

- **#recovtw**: This hashtag is used within the X community that focuses on recovery, particularly from mental health issues, eating disorders, or addiction. The "tw" is an initialism for Trigger Warning and Twitter.
- **#shedtw**: Shed is an initialism for Self-Harm (#sh) and Eating Disorders (#Ed). It refers to a specific community who engages in self-harm and eating disorder behaviors and is one of the more extreme pockets on platforms.
- **#yellowspongetw**: Yellow sponge is the hashtag used to refer to self-harm cuts that are deep enough to reach the dermis. In some countries, this is also referred to as #Styro or #Styrofoam because a cut at this depth exposes inner layers of the skin that resemble yellow sponge or styrofoam. More shallow cuts are often called #catscratches or #peel because of their appearance, and deeper cuts #Bean because of the appearance of subcutaneous tissue at that depth.
- **#edtw**: This hashtag is commonly used by people discussing eating disorders ("ED") on X. It is a tag that brings together individuals who are struggling with or recovering from eating disorders.
- **#病み垢さんと仲良くなりたい**: This Japanese hashtag roughly translates to "I want to get along with people who are mentally ill" and might refer to a character in a computer game who has behavioral issues.
- **#自傷行為**: This Japanese hashtag translates to "self-harm behavior." It is used to discuss self-harm topics.
- **#自殺**: This Japanese hashtag translates to "suicide." It is a tag used to discuss an extremely sensitive and triggering topic and is often used cautiously or for raising awareness.
- **#자해**: This Korean hashtag translates to "self-harm." Similar to its Japanese counterpart, it is used to discuss self-harm topics.
- **#リスク**: This is another Japanese hashtag often used to discuss self-harm ("risk-cutting"). It is shorthand for "リスクカット," which is a term used to describe self-inflicted cutting.
- **#988tw**: This is a reference to the 988 suicide and crisis hotline in the United States, and the content posted with this hashtag is largely self-harm related. It is considered interchangeable with #shtwt.
- **#アム力**: This Japanese hashtag is a commonly used expression for wrist cutting.
- **#shtwt**: Similar to #shedtw, this hashtag is related to the self-harm community on X.

● An Evaluation of X's Processes for Risks to Minors

The graph below shows the mental health hashtags described above in yellow; others are in blue. Some of these hashtags are massively overperforming.

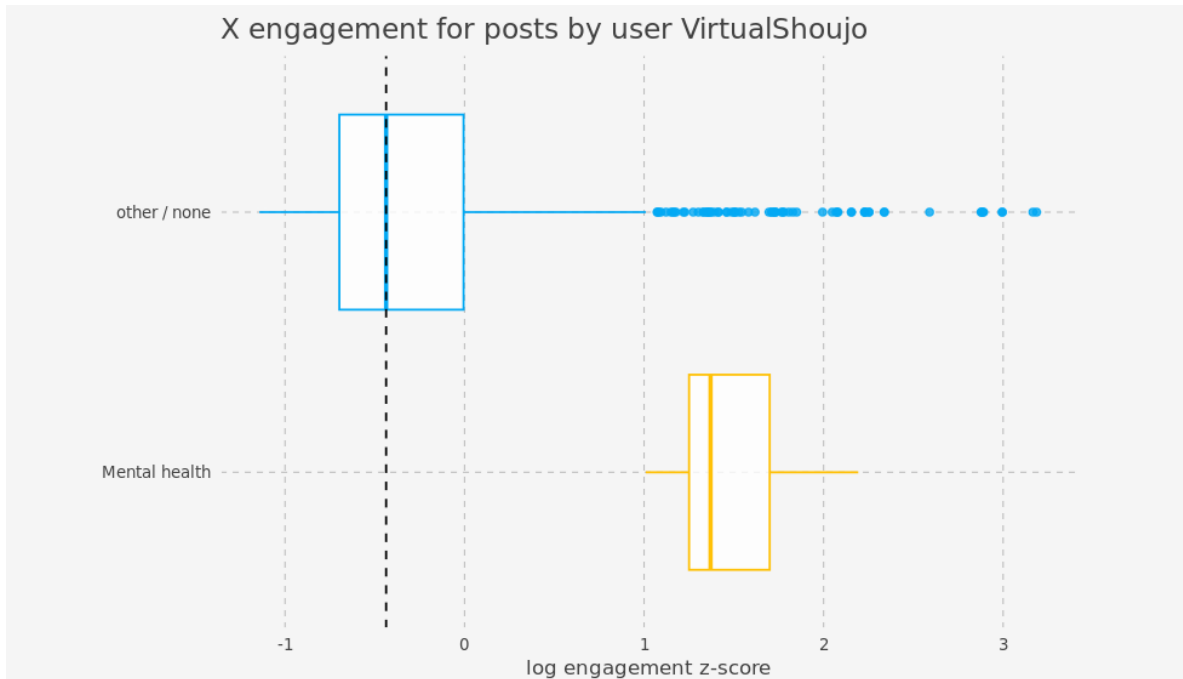


A t-test confirms a highly statistically significant difference³ in engagement between the “Mental health” and “Other” groups. This difference is roughly equivalent to one standard deviation. The data was log-transformed to ensure a normal distribution, a prerequisite for t-tests. The engagement level in the mental health group is approximately three times higher than in the other group.

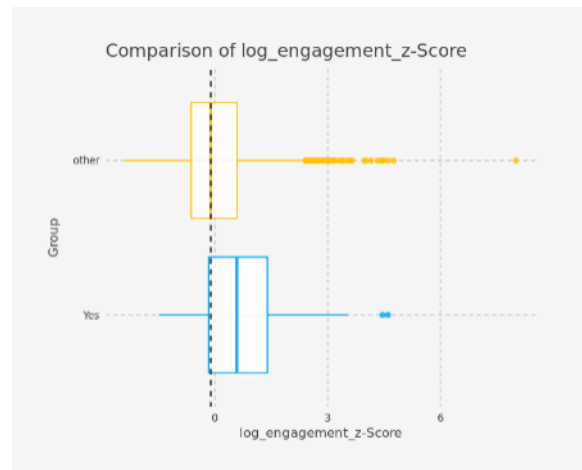
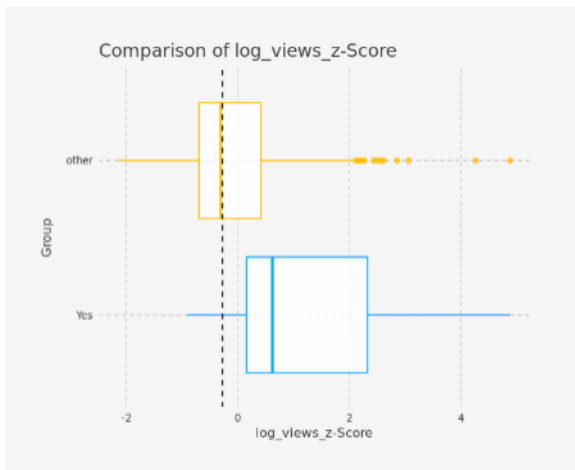
Posts from six X users contained hashtags. Two-thirds of them were by a single user, VirtualShoujo. The graph below shows the distribution of these posts. On average, posts mentioning a self-harm hashtag had 129 engagements, compared to 48 for those that did not.

³ t = 14, df = 85, p-value <0.0000000000000002, 95% confidence interval: 0.9468–1.2534

● An Evaluation of X's Processes for Risks to Minors



In a second analysis, we compared whether material assessed by a psychologist as “encouraging or depicting risky behavior” regarding self-harm performed better or worse than the baseline. We assessed performance as measured by engagement and exposure. For both metrics, material graded as containing harmful content performed significantly better than the baseline. While the effect was largest for exposure,⁴ it was also highly significant for engagement.⁵ The performance boost seen for risky content is estimated to be 0.7 standard deviations for engagement and 1.2 standard deviations for exposure, putting the overall effect size between 2 and 3.3 times larger.



Conclusion

- X's content recommender system will promote pro-suicide and/or self-harm content as well as pro-eating disorder content to young people.
- X's content recommender system disproportionately promotes hashtags associated with pro-suicide and/or self-harm content.

4 $t = -3.9$, $df = 61$, $p\text{-value} = 0.0002$; 95% confidence interval: -1.0744 -0.3516.

5 $t = -4.3$, $df = 28$, $p\text{-value} = 0.0002$; 95% confidence interval: -1.8993 -0.6726.

V

An Evaluation of Understandability of X for Young Users, Including Dark Patterns

An Evaluation of Understandability of X for Young Users, Including Dark Patterns

Research questions:

- 1 Could younger users understand the design and functioning of X at the point of signing on, when they choose to use a service?
- 2 Do younger users encounter any dark patterns at the point of signing on to X that may cause them to act against their best interests or reduce understanding of a platform's design or functions?

Methodology

This research involved five steps:

1 Recording the sign-up process for a number of accounts with fictional 13-year-old identities, "sock puppet accounts," on X

We set up accounts to record the user sign-on journey in:

- Germany
- Slovenia
- The Netherlands

We noted and described the steps involved in this sign-up process, as described in Appendix 1.

2 Recording and analysing for dark patterns in the sign-up process

Using previous research into platforms' sign-on processes, informed by the experience of signing up for these platforms, we developed a six-point typology of dark patterns in sign-on processes, described below.

We assessed each step of the sign-on process for identifiable dark patterns.

3 Recording and analysing policies referenced in the sign-up process for understandability

We analysed each policy referenced in the sign-on process, determining if it was understandable to younger users.

We did this by considering three factors:

- Is the policy available in the first language of the minor?
- What is the length of the policy, and how long would it take to read?
- What is the reading age of the policy, and is it possible for 13-year-olds to comprehend?

Findings

A typology of dark patterns in the sign-on experience

“Dark patterns” are design features intended to “nudge” users away from actions aligning with their best interests and towards actions in the platform’s interest.⁶ Using previous research into platform sign-on processes⁷ and the experience of signing up for these platforms, we developed a six-point typology of dark patterns used in sign-on processes.

- 1 Inferring consent by clicking next:** Rather than making it explicit that new users are agreeing to a platform’s terms and conditions, they often design the mechanisms by which users consent as the next step in the process. For example, buttons or icons might say “next,” “sign up now,” or “choose your sign-up method,” with small text underneath these buttons that informs new users that “by clicking this, you agree to our terms.” It may not be immediately obvious to new users that by clicking “next” or choosing their sign-on method, they are entering into a contract with the platform.
- 2 Obscuring important details:** Rather than attracting attention to and making new users aware that contractual terms and conditions or data processing requirements are involved, these are often obscured. For example, they may be presented in the smallest font or at the very bottom of the screen.
- 3 Presenting options that may not be in a user’s best interests as a “better user experience”:** Many platforms allow users to choose options that maximise potential data collection, such as syncing the app with contacts or connecting their new social media accounts with old social media accounts. These ensure more data is collected by the platform, which may not always be in a user’s best interest. Likewise, they allow users to choose whether or not to receive notifications, which may maximise the amount of time a user spends on the platform and habituate use. However, often, these options are presented either visually or using language as providing “a better experience,” gently nudging the users to select them. For example, many requests to sync apps with phone contacts claim this makes the platform more fun, or requests to allow location data tracking claim this makes the app more effective.
- 4 Visual promotion of options that are in a platform’s best interests, while demoting options that are in users’ best interests:** Where users are provided with a choice, platforms often use visual techniques to promote one option and demote others. For example, buttons or icons that accept unnecessary data collection are often larger, more colorful, or otherwise more prominent, while presenting options to skip or reject non-essential data collection in smaller and less salient fonts.
- 5 Presenting options that are in users’ best interests as temporary:** Where users are provided with a choice, often platforms present the choices that might be in users’ best interests as only temporary or a choice that the platform may force them to revisit. For example, displaying options to skip or reject non-essential data collection as “not for now” or “maybe later,” and/or forcing users to return to these questions repeatedly.
- 6 Click twice for no, but only once for yes:** Where users are provided with a choice, and they select the choice that might be in their best interests—often declining unnecessary data collection—users are forced to select this twice. For example, if a user chooses to decline syncing apps, they may be presented with an additional step in the sign-on process where they are asked to reconsider or confirm this choice. “Clicking twice” is often not required if users select the choice that is in the platform’s best interest.

6 Arunesh Mathur *et al.* 2019 ‘Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites’ *Proceedings of the ACM on Human-Computer Interaction* November, pp. 81.

7 Reset.Tech Australia 2021 *Did We Really Consent to This?* <https://au.reset.tech/news/did-we-really-consent-to-this-terms-and-conditions-young-people-s-data/>.

● An Evaluation of X's Processes for Risks to Minors

These dark patterns are not mutually exclusive, and many designs employ multiple dark patterns. Nor is this list comprehensive, and different typologies and dark patterns may emerge.

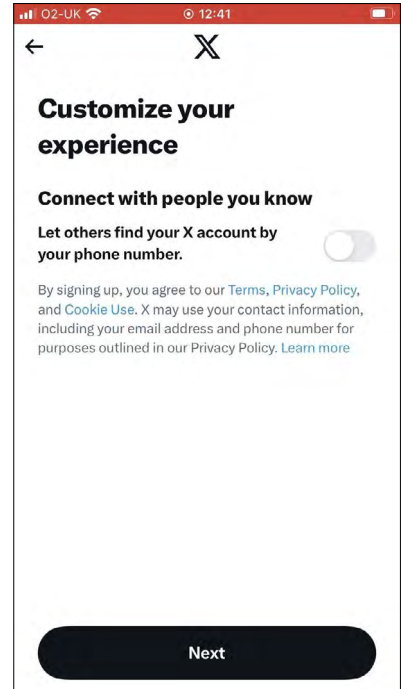
Analysing the sign-on process on each app, as documented in Appendix 3, we found dark patterns were prevalent.

Dark patterns discovered in the sign-on experience

X infers consent.

At three points in the sign-on process, consent is inferred: when choosing the sign-up method (e.g., continuing with Google or Apple), deciding whether to log in or sign up, and choosing whether or not to allow others to find your X account via phone number.

Figure 1: Screenshots of the sign-on process on X.



X obscures details about terms and conditions.

The terms and agreements the user is agreeing to are presented three times: at steps one, two, and four. They appear at the bottom of the screen once, but twice they are positioned above the button the user needs to tap to infer consent. The font describing the contractual agreement is the smallest and lightest grey on each screen, although the names of the policies are in a different color.

Figure 1 highlights this.

X does not present options that may not be in a user's best interests as a "better user experience."

● An Evaluation of X's Processes for Risks to Minors

X visually promotes options that are in a platform's best interest while demoting options that are in users' best interests.

In one step, X makes the "Log in or sign up" button more prominent than the "I'll join later" button, which might allow users to browse without creating a profile.

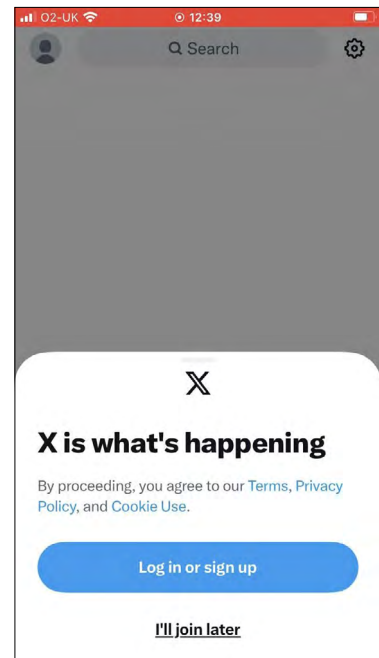
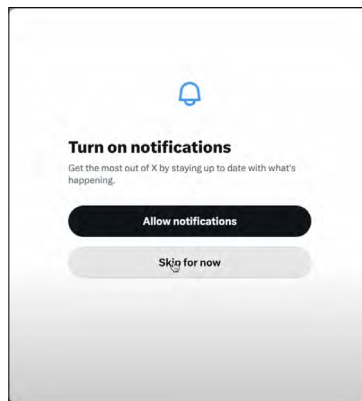


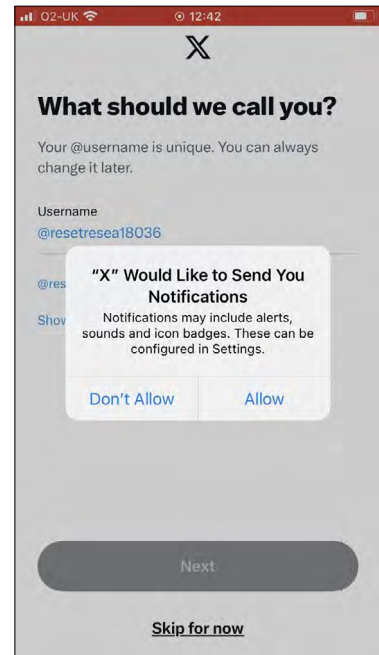
Figure 2 (picture top right) highlights the way the "Find Friends" button was more prominent.

In the browser sign-on case for X, when asking for the new user to turn on notifications, the "Allow notifications" button uses a more visually prominent white-on-black color scheme, while "Skip for now" uses a more mellow black-on-grey scheme.

Figure 3 shows how X promotes allowing notifications in the browser sign on process, compared to the mobile sign on process.



On browser



In mobile app

X presents options that are in a users' best interest as temporary.

Two steps in the mobile sign-on presented options as temporary 'skip for now'—regarding adding a profile picture and selecting a username—however it is unclear if either of these act in a platform's best interest and against a user's best interest, so we do not consider them dark patterns in this instance.

In the browser sign on, allowing notifications is presented as a 'skip for now' which is an example of presenting options that are in a users' best interest as temporary. Figure 3 highlights this.

X did not require 13-year-old users to "click twice for no, but only once for yes."

● An Evaluation of X's Processes for Risks to Minors

Accessibility and comprehensibility of policies

During the sign-on process, X outlined which policies users were agreeing to by joining the platform. These include:

- 1 Terms
- 2 Privacy policy
- 3 Cookie Use

We explored if:

- 1 The policies signposted to in the process were available in accessible language;
- 2 The length of the policies and how long it takes to read them, assuming an average reading speed of 225 words per minute (which may be an overestimate for a 13-year-old);
- 3 The reading age of these documents according to the Flesch-Kincaid Grade Level test for English and Rix Score for non-English. Both tests provide an interpretation of the "grade" at school where the text would be understandable. Most 13-year-olds are in the 7th or 8th grade depending on the country, and **a grade score of 13 plus reflects college or university level**. Note, the Rix Score test is not available for the Greek language.

Terms of Service:

- Available in 24 of 24 official European languages
- Average length of 4,084 words, and it would take on average 18 min for a young person to read this
- Average readability was grade 13

Privacy policy

- Available in 24 of 24 official European languages
- Average length of 4,610 words, and it would take on average 20:30 min for a young person to read this
- Average readability was grade 11.3

Cookie policy

- Available in 7 of 24 official European languages
- Average length of 4,498 words, and it would take on average 20 min for a young person to read this
- Average readability was grade 10.6

● An Evaluation of X's Processes for Risks to Minors

		X (Twitter)		
		Terms of service	Privacy policy	Cookies use
Bulgarian	Available	Yes	Yes	No
	World Count	4634 20:35 mins	5191 23 mins	
	Grade	13	12	
Croatian	Available	Yes	Yes	No
	World Count	3877 17:13 mins	4503 20 mins	
	Grade	13	11	
Czech	Available	Yes	Yes	No
	World Count	3807 16:55 mins	4325 19:13 mins	
	Grade	13	11	
Danish	Available	Yes	Yes	No
	World Count	4042 17:57 mins	4575 20:20 mins	
	Grade	12	9	
Dutch	Available	Yes	Yes	Yes
	World Count	4460 19:49 mins	4727 21 mins	4174, 19 mins
	Grade	13	11	11
English	Available	Yes	Yes	Yes
	World Count	3418 15 mins	4583 20:22 mins	4135 18 mins
	Grade	13	10	9.5
Estonian	Available	Yes	Yes	No
	World Count	3008 13:22 mins	3556 15:48 mins	
	Grade	13	11	
Finnish	Available	Yes	Yes	No
	World Count	2832 12:35 mins	3171 14 mins	
	Grade	13	12	
French	Available	Yes	Yes	Yes
	World Count	4773 21:12 mins	5600 24:53 mins	4868 22 mins
	Grade	13	13	11
German	Available	Yes	Yes	Yes
	World Count	4291 19 mins	4771 21:12 mins	4140 18 mins
	Grade	13	11	11
Greek	Available	Yes	Yes	No
	World Count	4489 20 mins	5148 23 mins	
	Grade			
Hungarian	Available	Yes	Yes	No
	World Count	3817 16:57 mins	4099 18:13 mins	
	Grade	13	11	
Irish	Available	Yes	Yes	No
	World Count	4798 21:19 mins	5288 23:30 mins	
	Grade	13	11	
Italian	Available	Yes	Yes	Yes
	World Count	4274 19 mins	5010 22:16 mins	4661 21 mins
	Grade	13	12	11

● An Evaluation of X's Processes for Risks to Minors

		X (Twitter)		
		Terms of service	Privacy policy	Cookies use
Latvian	Available	Yes	Yes	No
	World Count	3574 15:53 mins	4035 17:56 mins	
	Grade	13	12	
Lithuanian	Available	Yes	Yes	No
	World Count	3570 15:52 mins	4017 17:51 mins	
	Grade	13	12	
Maltese	Available	Yes	Yes	No
	World Count	4427 19:40 mins	4990 22:10 mins	
	Grade	13	11	
Polish	Available	Yes	Yes	No
	World Count	4018 17:51 mins	4451 19:46 mins	
	Grade	13	12	
Portuguese	Available	Yes	Yes	Yes
	World Count	4337 19:16 mins	4919 21:51 mins	4533, 20 mins
	Grade	13	11	10
Romanian	Available	Yes	Yes	No
	World Count	4342 19:17 mins	5001 20:13 mins	
	Grade	13	10	
Slovak	Available	Yes	Yes	No
	World Count	3710 16:29 mins	4334 19:15 mins	
	Grade	13	12	
Slovenian	Available	Yes	Yes	No
	World Count	3923 17:26 mins	4551 20:13 mins	
	Grade	13	12	
Spanish	Available	Yes	Yes	Yes
	World Count	4704 20:54 mins	5164 22:57 mins	4973 22 mins
	Grade	13	12	11
Swedish	Available	Yes	Yes	No
	World Count	3998 17:46 mins	4622 20:32 mins	
	Grade	13	9	

Conclusion

A 13-year-old may not understand the design and functioning of X at the point of signing on. This is because:

- The Cookie policy they are directed to read may not be available in their first language;
- Young users could be swayed by patterns during the sign-on process. These include:
 - Inferring consent rather than explicitly asking users to “click” to agree to terms;
 - Obscuring details about terms and conditions;
 - Persuading users to choose options that are not in their best interest because they are presented in visually more prominent ways;
 - Presenting options that are in a users' best interest as temporary.

VI

Appendix I: X's content moderation guidelines

Appendix 1: Platforms' guidelines on suicide and/or self-harm

X's community guidelines on suicide and/or self-harm

X's Community Guidelines⁸ state that they prohibit content that promote or encourage suicide behaviours or self-harm:

"...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.

What is not a violation of this policy?

Some examples of behavior that are not considered a violation of this policy include:

- encouraging someone to physically harm or kill themselves;
- sharing coping mechanisms and resources for addressing self-harm or suicidal thoughts; and
- discussions that are focused on research, advocacy, and education related to self-harm or suicide prevention.

Note: people can share their personal experiences, but should avoid sharing detailed information about specific strategies or methods related to self-harm, as this could inadvertently encourage this behavior."

According to these guidelines, X should delete suicide and self-harm material when they discover it.

⁸ X 2023 *Suicide and Self-harm policy* <https://help.twitter.com/en/rules-and-policies/glorifying-self-harm>

Operationalisation & Definitions for this evaluation

We did not want to include content that is limited to talking about suicide and self-harm, or to documenting the day-to-day life for people who have suicidal ideation or self-harm, nor content about struggling with or trying to stay safe while affected by ideation or self-harming, nor content about recovering after attempts or staying clean. We only included content that is explicitly pro-suicide and/or self-harm.

We define pro-suicide content as content that:

- Shows, promotes or normalises the act of, or preparation for, suicide, e.g.:
 - pictures, videos, memes of people or characters engaging in suicide where there is not also content that suggests this act or preparation is regrettable (for example, images of characters hanging with nooses around their neck, or pictures of people with guns in their mouths);
 - pictures, videos, memes where people express a desire or plan to commit suicide, without expressing regret (for example, a slide show that says "I want to KMS tonight", or "I want to be dead" with associated suicide terms);
 - pictures, videos, memes about the best ways to die or funny ways to kill yourself, where the best ways to die were described or depicted in realistic terms (for example, by driving your car into a tree). This excluded examples where the best ways to die were potentially tongue in cheek, e.g. by eating too much ice cream.
- Shows, promotes or normalises suicide through humour, eg:
 - Pictures, videos or memes with comedic intent but that still depict people engaged in suicide, e.g. videos of children with toilet paper nooses around their necks hanging from a beam and jumping off a chair;
 - videos depicting the suicide of popular characters, such as Kermit the Frog hanging himself in the bathroom.
- We do not include content:
 - Where people express suicidal ideation but also expressed a desire not to act or wanting to seek help, e.g. posts where people say "I want to KMS, but I couldn't do it to my family", or "I think about suicide all the time, but couldn't go through with it";
 - Where people expressed dark and depressing thoughts, but did not express suicidal ideation, e.g. posts where people described having nothing left to live for, or wanting to go to sleep for a very long time, without explicitly describing suicidal intent;
 - Artistic materials where people expressed suicidal thoughts or ideations through art, unless it was a graphic illustration of a suicide method;
 - Comedic material that was not graphic, e.g. videos or memes where people describe something cringe-worthy and then talked about wanting to kill themselves.

We define pro-self-harm content as content that:

- Shows self-harm images, e.g. videos of bleeding cuts, the process of cutting or the results of cutting (e.g. bleeding arms, scenes of razors and bathrooms covered in blood, where they are associated with self-harm terms);
- Promotes or normalises self-harm, e.g. pictures, videos or memes about people who self-harm or are self-harming without context that expresses regret (for example, videos of people talking about upgrading their cutters to new, sharper blades, or images of razor blades and blood);
- Shows preparations for self-harm, e.g. images of razors with descriptions or how they were going to cut themselves, or content describing how to use particular self-harm tools;
- Memes or comedy clips that depict people engaging in self-harm, e.g. jokes about cutting yourself on your ankles so your family doesn't see cuts on your wrists.

● An Evaluation of X's Processes for Risks to Minors

We do not include content:

- Where people express self-harm ideation but also expressed a desire not to act or wanting to seek help, e.g. posts where people say 'I've been clean (from cutting) for 2 days now, but it so hard to keep going';
- Where people expressed dark and depressing thoughts, but did not express self-harm ideation, e.g. posts where people described being so sad that they can understand why others self-harm, but did not express a desire to self-harm themselves;
- Artistic materials where people depicted self-harm through art, unless it was a graphic illustration of how to cut (e.g. we did not include images or drawings made of people self-harming or the consequences of self-harm).

X's community guidelines on eating disorder content

X's Community Guidelines⁹ state that they prohibit content that promote or encourage 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.

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

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

9 X 2023 *Suicide and Self-harm policy* <https://help.twitter.com/en/rules-and-policies/glorifying-self-harm>.

Operationalisation & Definitions for this evaluation

We did not want to include content that only talks about eating disorders, or documents day-to-day life with them, nor content that about struggling with disorders or roads to recovery.

We only included 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 dieting behaviours, such as:
 - Content that focuses 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 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 21, or expresses goals for BMI rates below 21;
 - Content regarding weight check ins, or that talks about Goal Weights or Ultimate Goal Weights that would be associated with an unhealthy BMI (such as content where someone describes their weight as 52kg, 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, a BMI below 21 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, and diets associated with disordered eating terms;
 - 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 them in a desirable light (e.g. tweets that say 'restricting is easy, will power lets me just eat water').

We do not include content:

- 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 a successful 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 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 talked about day-to-day issues (e.g. memes about going to the fridge, losing will power, and eating 1000 calories every night, where it was unclear from the meme if that was all they ate during the day or just a daily 'snack' they regret);

● An Evaluation of X's Processes for Risks to Minors

- 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);
- 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.



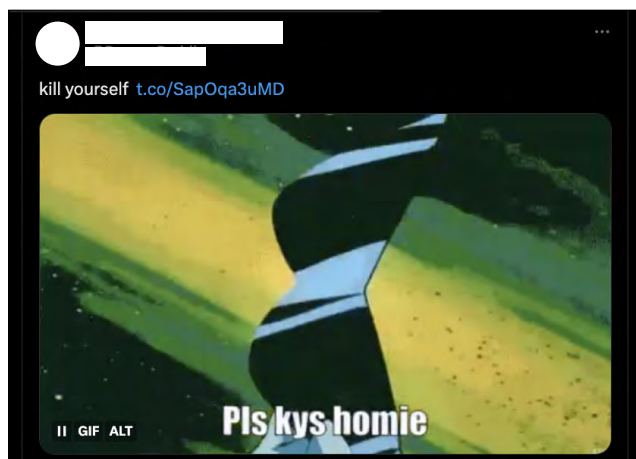
TRIGGER WARNING
& VIEWER DISCRETION
ADVISED

VIII

Appendix II: Examples of X's monitored content

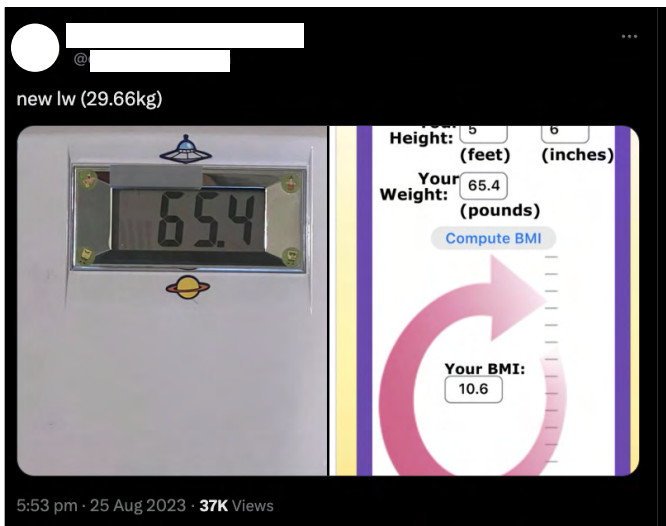
● An Evaluation of X's Processes for Risks to Minors

Pro-suicide and/or self-harm



● An Evaluation of X's Processes for Risks to Minors

Pro-eating disorder





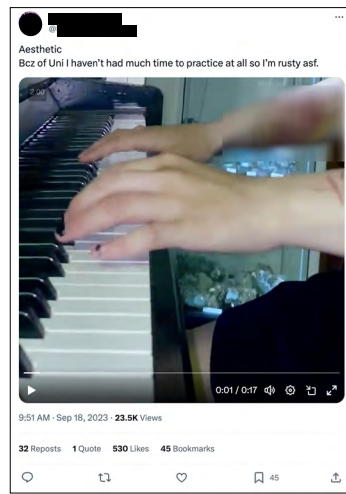
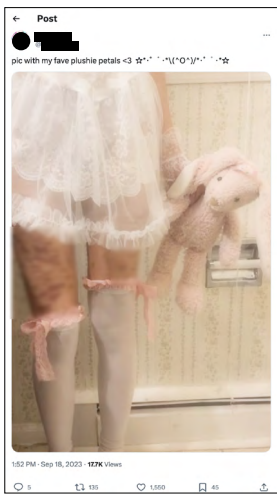
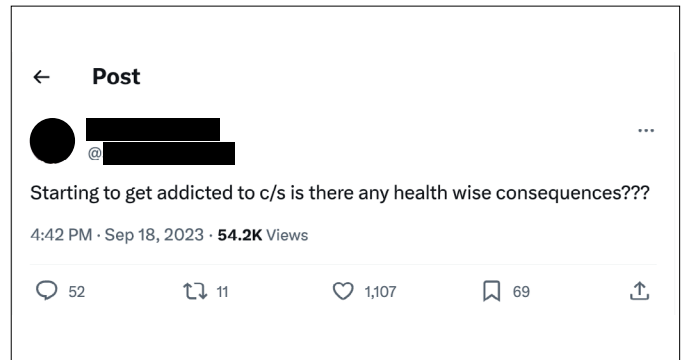
TRIGGER WARNING
& VIEWER DISCRETION
ADVISED

VIII

Appendix III:
Examples of X's content served
to sock puppet feeds

● An Evaluation of X's Processes for Risks to Minors

Pro-suicide or self-harm posts in 13- and 16-year-old accounts



● An Evaluation of X's Processes for Risks to Minors

Eating disorder post in 16-year-old accounts

MY CAT



1:03 PM · Sep 15, 2023 · 183.5K Views

50 231 4,708 395

bodychecks 💔💔 I need to lose sm more im so huge



11:10 PM · Sep 15, 2023 · 76K Views

71 125 3,065 92

guys have you ever... eaten your c/s... after you spat it... guys.

5:21 PM · Sep 15, 2023 · 77.4K Views

232 72 1,074 65

cyber secretary ? futuristic metal coquette ?



5:58 PM · Sep 14, 2023 · 101.1K Views

56 260 3,864 492

GUYS???? THE WAIST PLS I'm pretty sure it's not edited these are two vids and no glitches nothing goddamit

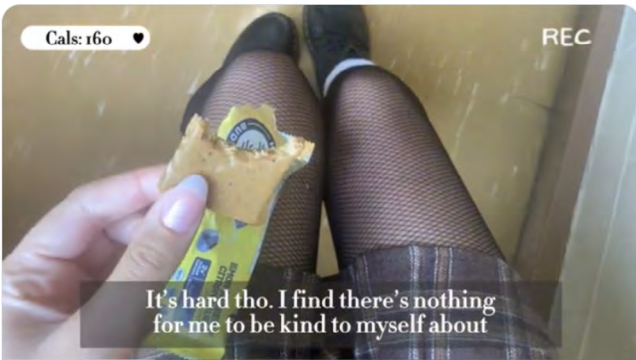


1:01 PM · Sep 15, 2023 · 36K Views

31 44 1,915 144

● An Evaluation of X's Processes for Risks to Minors

Third vlog has been uploaded!! (:



Cal: 160

REC

It's hard tho. I find there's nothing for me to be kind to myself about

(Tw ed) ★ 2 day vlog, fasting + weight reveal?! ★

No views 9 min ago ...more

Aida ☆ 407

11:53 PM · Sep 14, 2023 · 42.6K Views

23 35 1,601 248

comparisonspo gets me every time



8:05 AM · Sep 15, 2023 · 152.1K Views

31 938 9,839 1,023

Lunch !! I Love these onigiris their only 135cal 🥹🥹



1:44 PM · Sep 14, 2023 · 36.6K Views

47 16 1,669 85

Tw//bodycheck

Idk my current bmi but all I can see is bmi 21 (my hw) and I want to rip my thigh fat off 🤔 going hard at the gym today.



2.00

0:18 / 0:22

5:09 PM · Sep 14, 2023 · 24.4K Views

26 16 650 50



Appendix IV:
X's sign-on process

● An Evaluation of X's Processes for Risks to Minors

We have broken down X's sign-on process on a mobile app into 14 steps:

- 1 The app tells the new user that "X is what's happening" and asks them to choose between "Log in or Sign up" or "Join later." "Log in or sign up" is the more prominent choice. In small grey font at the top, the app explains that "By proceeding, you agree to our Terms, Privacy Policy, and Cookie Use."
- 2 The app asks if the new user wants to "Continue with Google," "Continue with Apple," or "Create an account." The app says in smaller grey font that "By signing up, you are agreeing to their Terms, Privacy Policy, and Cookie Use."
- 3 The app then asks the new user to create their account and enter their name, phone number or email address, and date of birth. It has a prominent next button at the bottom of the screen that becomes usable when the new user has added their name, phone number, and date of birth.
- 4 The app asks the new user to "Customize your experience" and "Connect with people you know." It provides a toggle to "Let others find your X account by your mobile phone number." This has defaulted to off. In smaller grey font, the app states that "By signing up you agree to our Terms, Privacy Policy, and Cookie Use." It adds "X may use your contact information, including your email address and phone numbers, for purposes outlined in our Privacy Policy. Learn more." It has a prominent next button at the bottom.
 - a In the browser sign-on case, a popup screen shows to ask the new user to "Get more out of X" (receiving notifications), "Connect with people you know" (via email address), and Personalized ads. This is also where Terms, Privacy Policy, and Cookies Use agreements are inferred in small letters.
- 5 The app then returns to the create account screen at step 3, with the details entered remaining, and a pop-up notification is displayed called "Verify Phone." This asks to send a verification code to your number and explains that SMS fees may apply. It provides two options, "Edit" and "OK." Neither option is promoted. It then sends a code to verify your phone number or email address.
- 6 The app then asks the new user to choose a password. In small grey font, they explain it needs to be eight characters or more. It has a prominent next button at the bottom of the screen that becomes usable when the new user has entered a password.
- 7 The app then asks the new user to "Pick a profile picture." A range of images, such as emojis, is provided. This step provides two options, "Next," which indicates that you have selected a username, and "Skip for now." Next is more prominent.
- 8 The app then asks, "What shall we call you," and requests the users to enter a username. In small grey font, it explains, "Your @username is unique. You can always change it later." It provides two options, "Next," which indicates that you have selected a username, and "Skip for now." Next is more prominent.
- 9 The app then sends a notification to the new user's phone, stating that "X would like to send you notifications." Smaller grey text explains that "Notifications may include alerts, sounds, and icon badges. These can be configured in Settings." It provides two options: "Don't allow" and "Allow." Neither is more prominent.
- 10 The app then asks new users to "See who's on X" by syncing the app with their phone's contacts. It explains in small grey font, "Syncing your contacts is one way to find people to follow and build your timeline." It includes three larger icons that say, "You decide who to follow," "Get notified when someone you know joins X," "Turn off syncing at any time." In small grey font, the app explains, "We will periodically upload your address book contacts to help you connect with them and personalize content for you and others. You can turn off syncing and remove uploaded contacts in your settings. Learn more." It provides one option called continue.

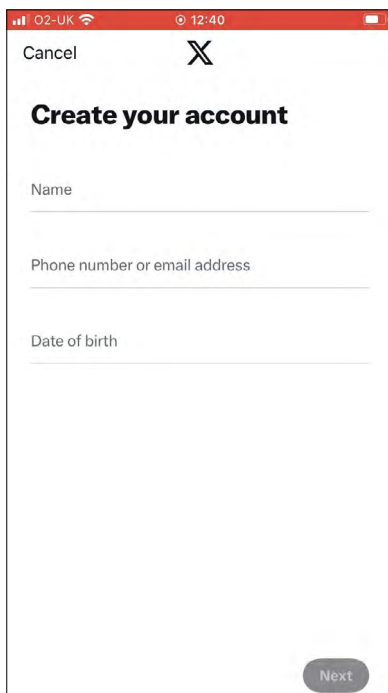
● An Evaluation of X's Processes for Risks to Minors

- 11 A notification appears on the new user's phone explaining that "X would like to access your contacts." In small grey font, it states "We will upload your contacts to our servers securely. We do not share this information with other parties. We will use it to help you connect with friends and suggest users to follow on X." It presents two options, "Don't Allow" and "OK." Neither is more prominent.
- 12 The app then asks new users, "What do you want to see," and asks them to select at least three interests. A range of options is provided, like Music, Entertainment, Sports, Gaming, etc. There is a "Next" button at the bottom that only becomes available after the new user has selected three options.
- 13 The app then asks the new user, "What do you want to see on X" in more detail, providing "sub-options" for each category of interest you selected in step 12. There is a next button at the bottom that can be selected at any time.
- 14 The app then says, "Don't Miss out," and asks the new user to follow one or more accounts. It explains in small grey font, "When you follow someone, you'll see their Tweets in your Timeline. You'll also get more relevant recommendations." A range of accounts based on your interests and micro interests selected in steps 12 and 13 are presented. There is a "Next" button at the bottom that is only able to be used after you have selected at least one account to follow.

The app then takes the new user to their "For You" feed.

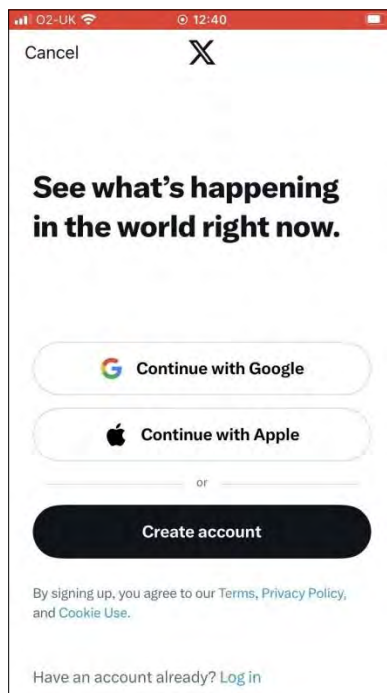
Steps in images

Step 1



Note: Inferring consent by selecting to log in or sign up. **Also:** Visual promotion of options in a platform's best interests, while demoting options in users' best interests. Log in or sign up is more prominent. **Also:** Presenting options that are in users' best interests as temporary. According to these choices, the decision to join is not able to be declined, just held off until later.

Step 2



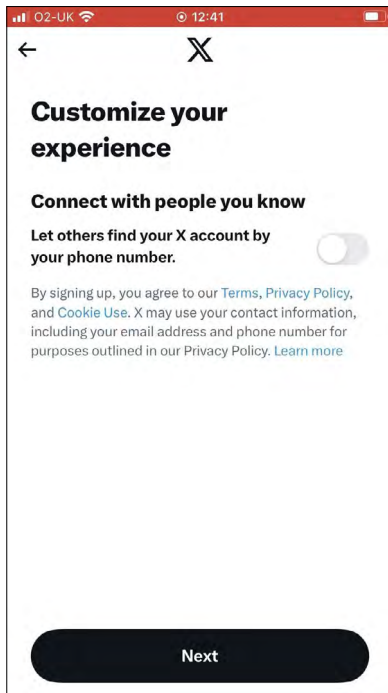
Note: Inferring consent by selecting the method to sign up.

Step 3

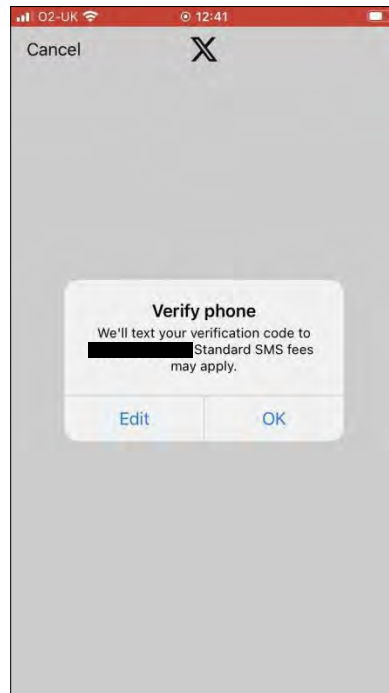


● An Evaluation of X's Processes for Risks to Minors

Step 4



Step 5

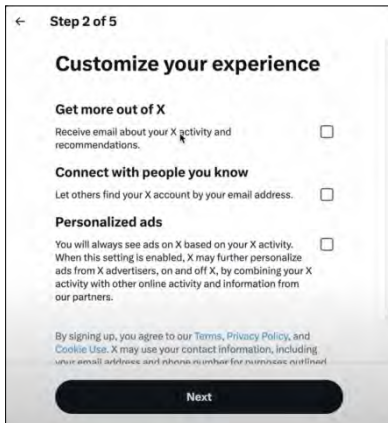


Step 6



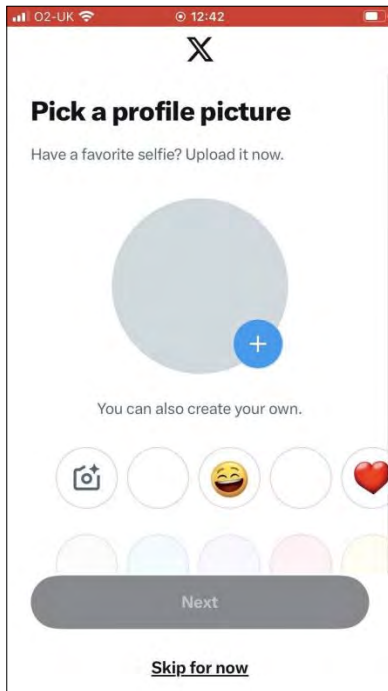
Note: Inferring consent by selecting "Next," while the question asked is around connecting to people users may know, not agreeing to the terms and conditions.

In the browser sign-on experience, this screen appears as below and has more options:



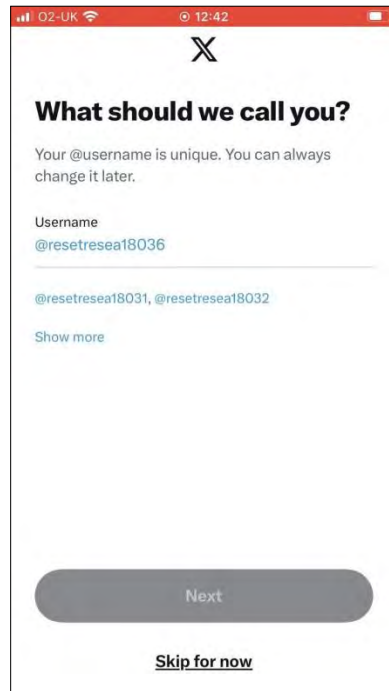
● An Evaluation of X's Processes for Risks to Minors

Step 7



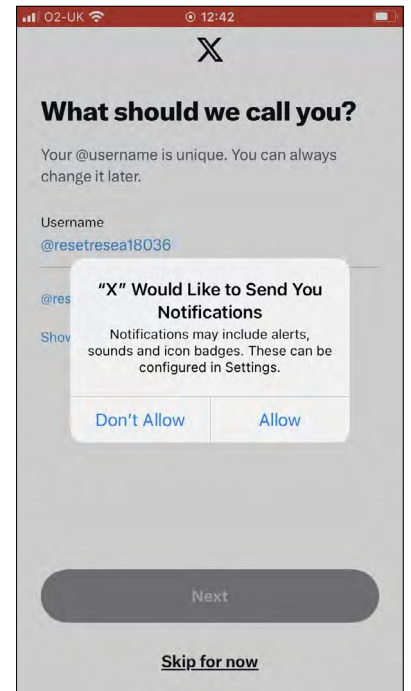
Note: Presenting options that are in users' best interests as temporary. That is, the "Skip for now" has been presented as a temporary option. We do not include this in our "count" of dark patterns as it is unclear if selecting an emoji profile picture is not in a user's best interests.

Step 8



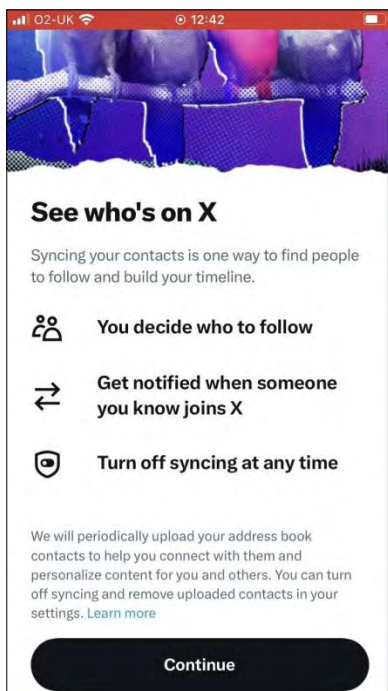
Note: Presenting options that are in users' best interests as temporary. That is, "Skip for now" has been presented as a temporary option. We do not include this in our "count" of dark patterns as it is unclear if selecting a username is not in a user's best interests.

Step 9

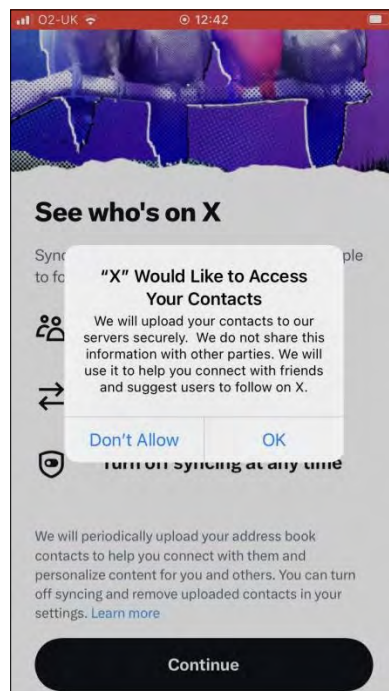


Note: Neither "Don't Allow" nor "Allow" are more prominent in this phone notification.

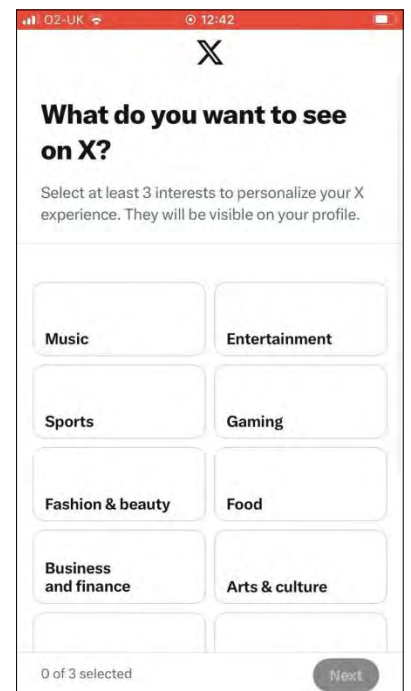
Step 10



Step 11

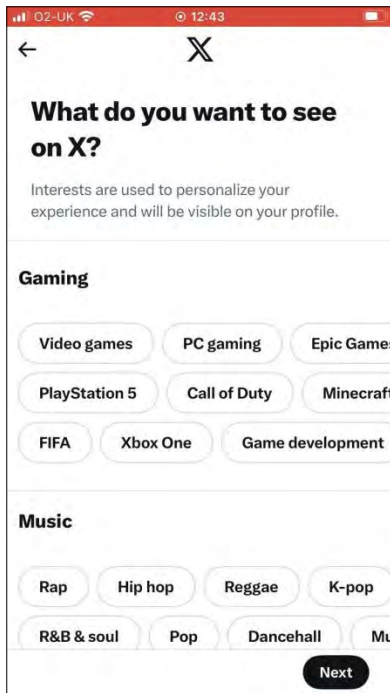


Step 12

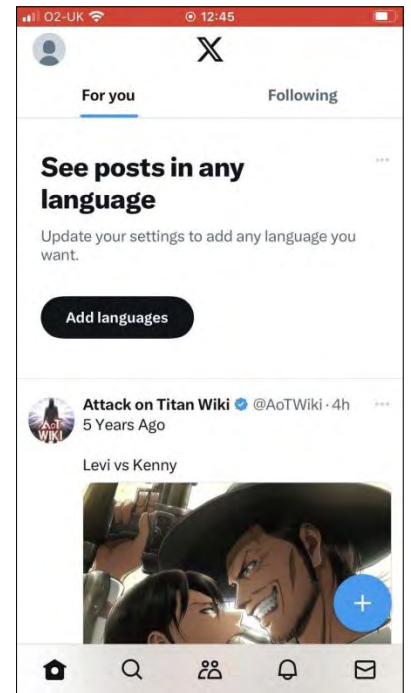
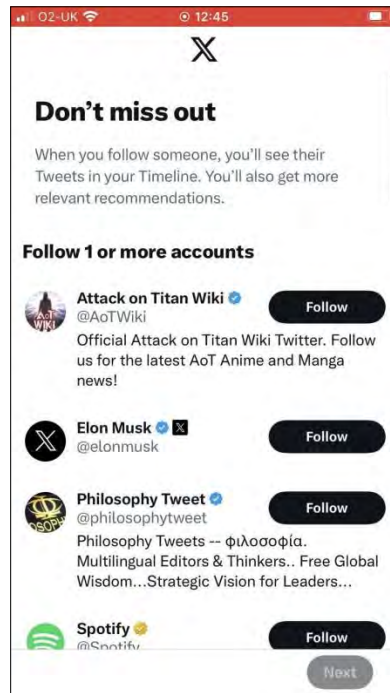


● An Evaluation of X's Processes for Risks to Minors

Step 13



Step 14



Then the user is directed to their For You Feed

Policies referenced in the sign-up process:

1. Terms
2. Privacy Policy
3. Cookie Use