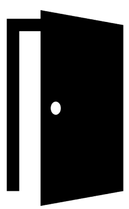
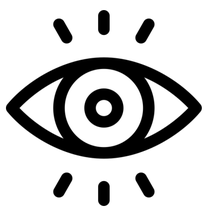
This is the ASPARC framework.

It serves to break down the user-generated content journey into 7 distinct phases.

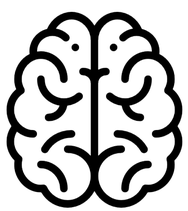
Access



Sign-on

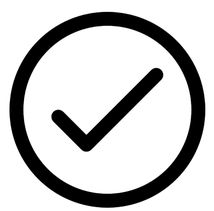


Participate



Analyse

Respond



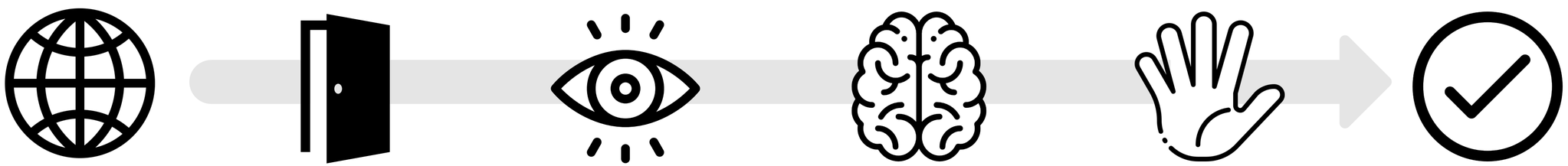
Comply

The models in this document were built around the ASPARC framework.

For each phase, from Sign-on to Comply, there is a functional model that provides a detailed view of the functional and architectural processes that platforms take to achieve the objectives called out below. 'Access', the process of accessing the internet, was out of scope for our investigation but is included below for completeness.

The models are generically representative of all online platforms, however significant areas of local diversity of implementation are highlighted. For more detail, please refer to the report.

Click on the icons to view the models.



Access

- Select device
- Select operating system
- Select Network access
- Select on-device safety features
- Deploy On-Network safety features

Sign-on

- Registration
- Establish user profile
- Age verification
- User sign in
- Connect with other users

Participate

- Create content
- Curate content
- Post static content
- Livestream content
- View content
- Report content

Analyse

- Train classifiers
- Auto-detect content
- Receive user signal
- Operate intelligence desk
- Auto-assess content against single classifiers
- Auto-assess content against combination classifiers
- Human moderation

Respond

- Relegate content (reduce virality)
- Remove content
- Sanction user
- Alert safety partners
- Alert law enforcement
- Invoke content incident protocol

Comply

- Manage appeals process
- Maintain MI database of harmful content and bad actors
- Share safety data with 3rd party partners
- Publish transparency report
- Audit performance

Ofcom

Phase 1

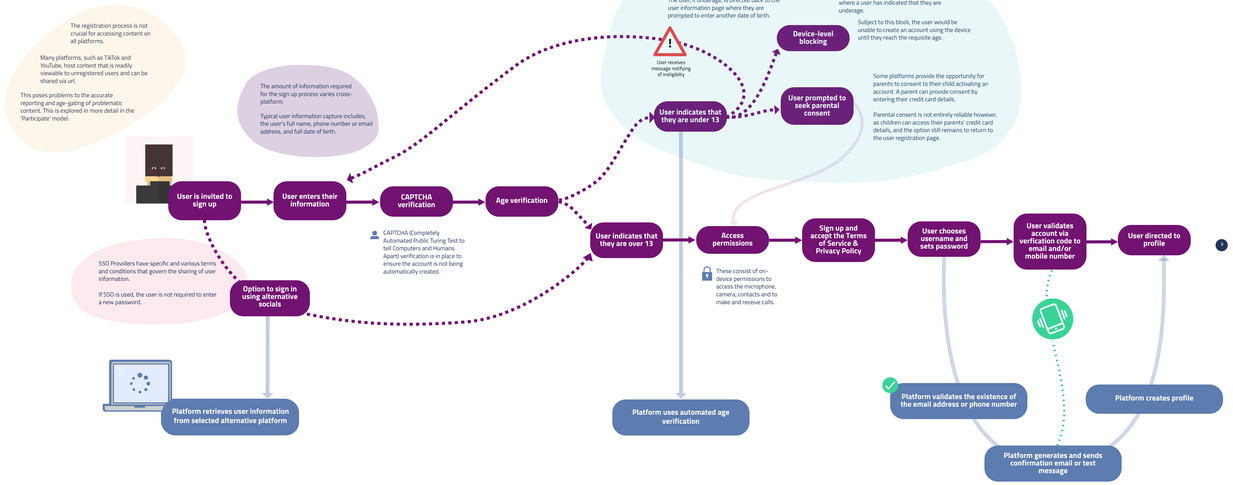
Sign-on

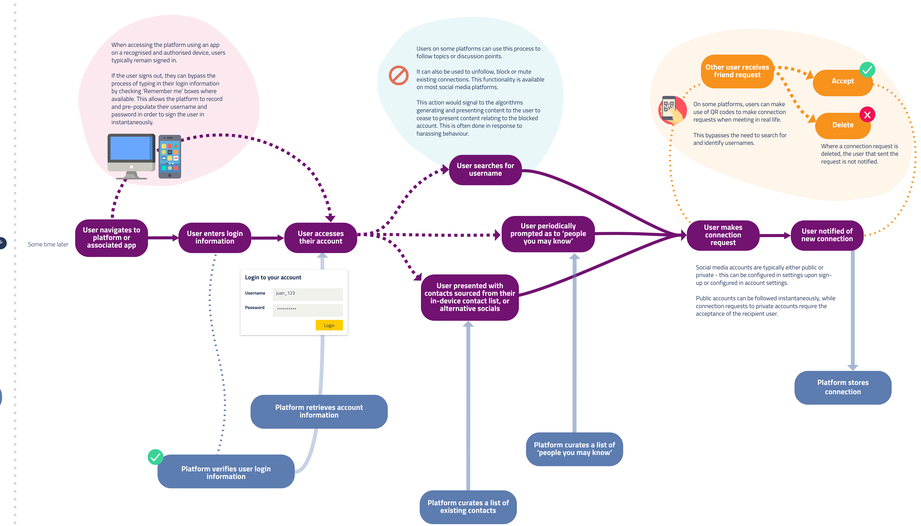
The 'Sign-on' model represents the process by which a user signs up and establishes their account. It covers user registration, age verification, sign in and the generation of connection requests for other users.

Key

Step in user journey

Platform activities





Architectural overlay

Ofcom

Phase 1

Sign-on

The 'Sign-on' model represents the process by which a user signs up and establishes their account. It covers user registration, age verification, sign in and the generation of connection requests for other users.

Key

Step in user journey

Platform activities

- security
- machine learning
- database/ persistence
- cloud
- external service
- search/analysis
- harmful content
- profile/registration
- image recognition
- management information capture

The registration process is not crucial for accessing content on all platforms.

Many platforms, such as TikTok and YouTube, host content that is readily viewable to unregistered users and can be shared via url.

This poses problems to the accurate reporting and age-gating of problematic content. This is explored in more detail in the 'Participate' model.

The amount of information required for the sign up process varies cross-platform.

Typical user information capture includes, the user's full name, phone number or email address, and full date of birth.

User is invited to sign up

User enters their information

CAPTCHA verification

CAPTCHA (Completely Automated Public Turing Test to tell Computers and Humans Apart) verification is in place to ensure the account is not being automatically created.

Age verification

In most cases this is the extent of the age verification functionality in place.

The user, if underage, is directed back to the user information page where they are prompted to enter another date of birth.

User receives message notifying of ineligibility

User indicates that they are under 13

Device-level blocking

Some platforms operate device-level blocks where a user has indicated that they are underage.

Subject to this block, the user would be unable to create an account using the device until they reach the requisite age.

User prompted to seek parental consent

Some platforms provide the opportunity for parents to consent to their child activating an account. A parent can provide consent by entering their credit card details.

Parental consent is not entirely reliable however, as children can access their parents' credit card details, and the option still remains to return to the user registration page.

User indicates that they are over 13

Access permissions

These consist of on-device permissions to access the microphone, camera, contacts and to make and receive calls.

Sign up and accept the Terms of Service & Privacy Policy

User chooses username and sets password

User validates account via verification code to email and/or mobile number

User directed to profile

Platform uses automated age verification

Natural language processing is used for text, PageRank for network analysis, and neural networks for image recognition.

Platform validates the existence of the email address or phone number

Platform generates and sends confirmation email or text message

Platform creates profile

User details stored on NoSQL or SQL databases hosted in-house or by a Cloud service provider; networks can be stored on Graph databases.

SSO Providers have specific and various terms and conditions that govern the sharing of user information.

If SSO is used, the user is not required to enter a new password.

Option to sign in using alternative socials

Platform retrieves user information from selected alternative platform

Authentication achieved using OpenID Connect and OAuth2. These are authorization frameworks that enable applications to obtain limited access to user accounts.

Where there is a business relationship between platforms, architectural components are shared.

Some time later

When accessing the platform using an app on a recognised and authorised device, users typically remain signed in.

If the user signs out, they can bypass the process of typing in their login information by checking 'Remember me' boxes where available. This allows the platform to record and pre-populate their username and password in order to sign the user in instantaneously.

User navigates to platform or associated app

User enters login information

Login to your account

Username: Juan_123

Password: ••••••••••

Login

User accesses their account

Platform retrieves account information

Platform verifies user login information

Users on some platforms can use this process to follow topics or discussion points.

It can also be used to unfollow, block or mute existing connections. This functionality is available on most social media platforms.

This action would signal to the algorithms generating and presenting content to the user to cease to present content relating to the blocked account. This is often done in response to harassing behaviour.

User searches for username

User periodically prompted as to 'people you may know'

User presented with contacts sourced from their in-device contact list, or alternative socials

Platform curates a list of existing contacts

Dependent on whether the user has given the app in-device permission to access contacts.

Platform curates a list of 'people you may know'

This list is curated on the basis of collected user information.

Depending on the platform, this can mean a user's historic activity (indicating interests), the groups a user participates in, or any further information that may indicate which accounts a user is likely to want to interact with.

Other user receives friend request

Accept

Delete

On some platforms, users can make use of QR codes to make connection requests when meeting in real life.

This bypasses the need to search for and identify usernames.

Where a connection request is deleted, the user that sent the request is not notified.

User makes connection request

User notified of new connection

Social media accounts are typically either public or private - this can be configured in settings upon sign-up or configured in account settings.

Public accounts can be followed instantaneously, while connection requests to private accounts require the acceptance of the recipient user.

Platform stores connection

The profile information will be indexed and there will be a primary key/ foreign key relationship with data relating to connections, usage and preferences.

This storage could be provided by a Cloud service provider.

Architectural overlay

Ofcom

Phase 2

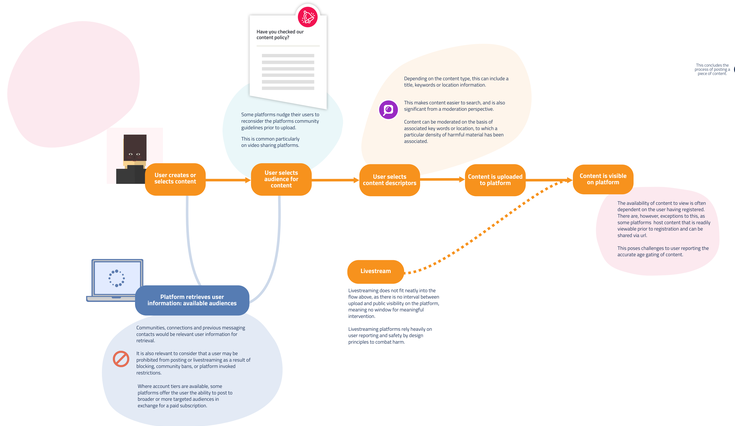
Participate

The 'Participate' model details a user's engagement with content on a platform. It covers the processes of creating, uploading and curating content, the various methods of discovering content to view and interact with, and the process by which users report objectionable content.

Key

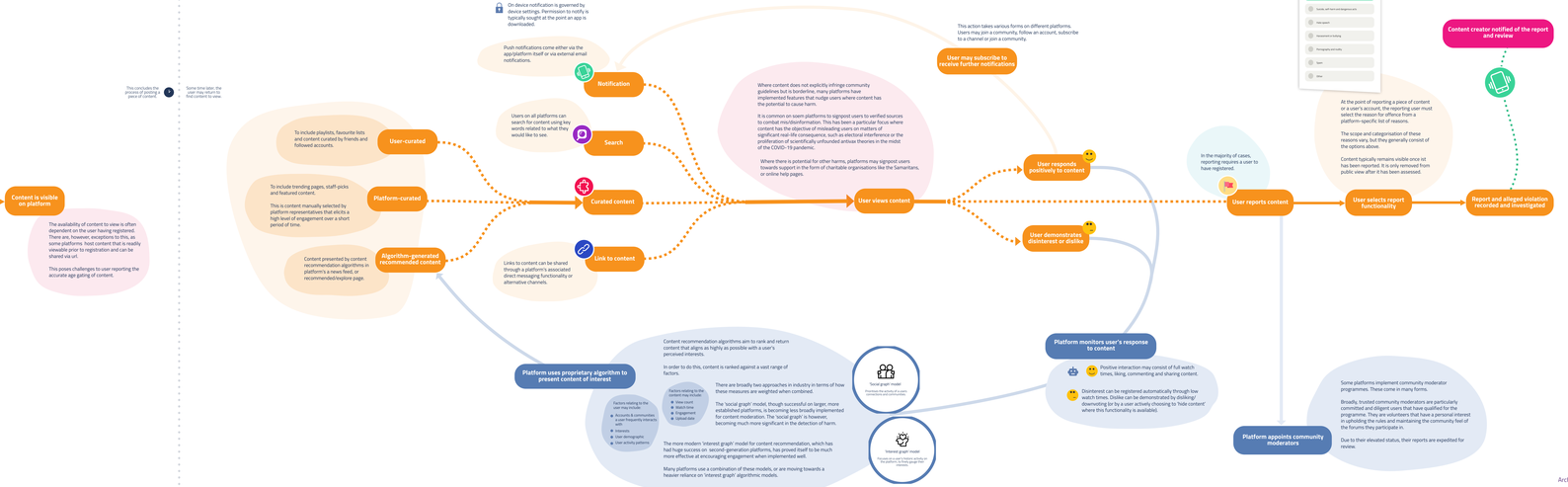
Step in user journey

Platform activities



This concludes the process of posting a piece of content.

Some time later, the user may return to find content to view.



Architectural overlay

Ofcom

Phase 2

Participate

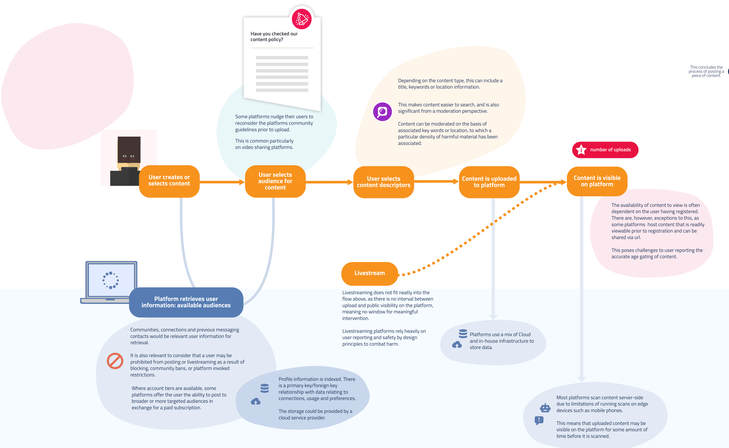
The 'Participate' model details a user's engagement with content on a platform. It covers the processes of creating, uploading and curating content, the various methods of discovering content to view and interact with, and the process by which users report objectionable content.

Key

Step in user journey

Platform activities

- security
- machine learning
- database/ persistence
- cloud
- external service
- search/analysis
- harmful content
- profile/registration
- image recognition
- management information capture



Architectural overlay

Ofcom

Phase 3

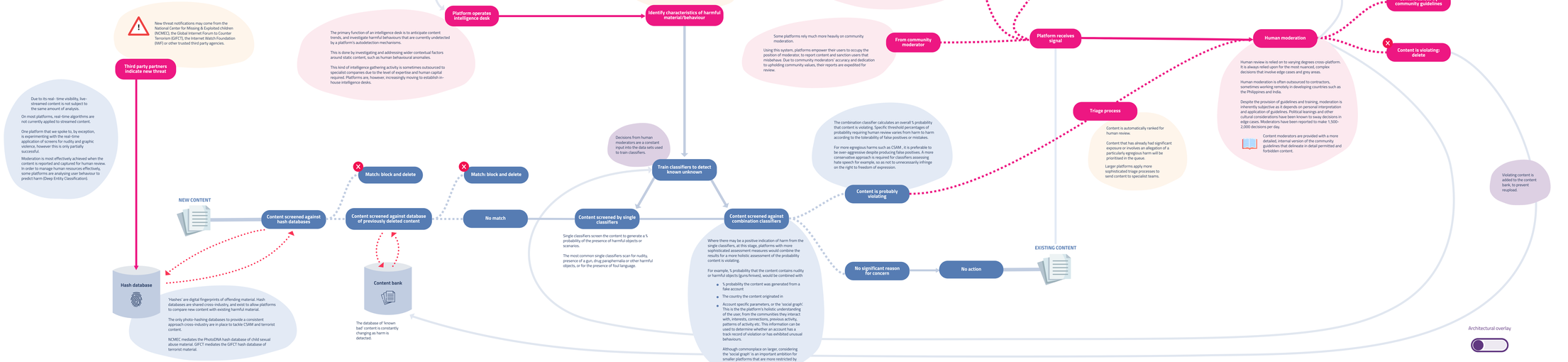
Analyse

The 'Analyse' model details the measures put in place to analyse content and user behaviour in order to determine whether it is harmful. These measures include the operation of an intelligence desk, the receipt of user signals, the application of single and combination classifiers and the triage process in place to prioritise content for human review.

Key

Human activity on behalf of platform

Platform activities



Ofcom

Phase 3

Analyse

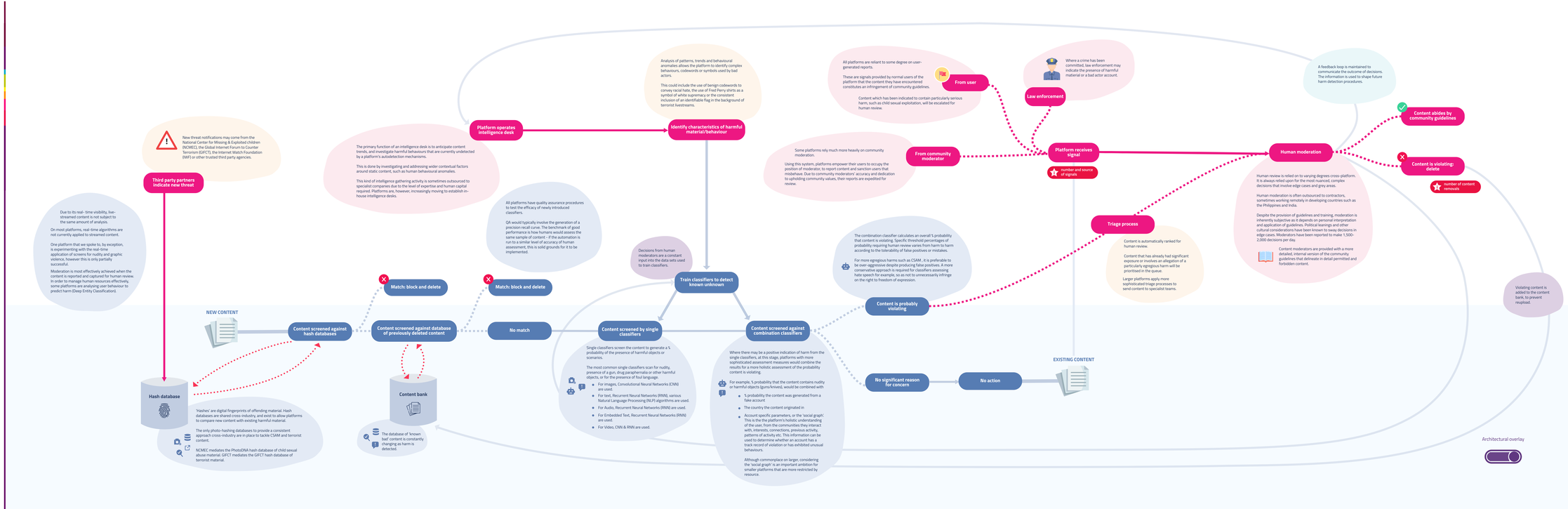
The 'Analyse' model details the measures put in place to analyse content and user behaviour in order to determine whether it is harmful. These measures include the operation of an intelligence desk, the receipt of user signals, the application of single and combination classifiers and the triage process in place to prioritise content for human review.

Key

Human activity on behalf of platform

Platform activities

- security
- machine learning
- database/ persistence
- cloud
- external service
- search/analysis
- harmful content
- profile/registration
- image recognition
- management information capture



Ofcom

Phase 4

Respond

The ‘Respond’ model provides detail on how platforms respond to the finding of harmful content. It covers the process of content removal, the application of user sanctions and the procedures in place to ensure that the relevant authorities are alerted.

Key

Step in user journey

Platform activities

User creates or shares harmful content

User notified of content removal

Sometimes user that originated the report is notified of the action a platform takes in response.

Some platforms have spoken of the benefits of maintaining this feedback loop. Where a user is able to see follow through and affirmative action as a result of signalling harm to the platform, they are more likely to do so again.

User responsible for the report notified of follow-through

Platform removes content

Where content is not a direct violation of a platform’s policies but is borderline, it will be relegated in content recommendation algorithms.

Some time later

Indicators of inauthenticity are commonly referred to as ‘co-ordinated inauthentic behaviour’ (CIB) in industry.

CIB can consist of liking too many things in too short a period, posting too much in too short a period, or having the same profile information as other accounts.

This behaviour points to the illegitimacy of the account and the interference of bots for the purposes of spam. CIB is the most common reason for account suspension.

Indication that the account is not legitimate

User violates community guidelines in a particularly egregious way

User repeatedly violates community guidelines

The definition of an egriegious violation and the threshold number of violations to require account suspension varies cross-platform.

On community-moderated platforms, the standard is defined by the user community.

Other platforms implement a ‘three-strike’ system, where a user receives two warnings for inappropriate behaviour before their account is suspended.

All platforms recognise the importance of consistent enforcement of user sanction.

The ultimate punitive measure platforms take is account deletion.

Depending on policy and their tolerance of bad behaviour howver, some platforms may implement temporary account suspensions. This would mean that the user subject to the measure would be unable to access their account for a specified period of time.

User notified of account deletion

Platform deletes user account

Account details are preserved to prevent ban evasion.

Platform outsources response

The Content Incident Protocol is a formalised playbook of actions formulated by the Global Internet Forum to Counter Terrorism (GIFCT).

It is the process by which a platform becomes aware of, quickly assesses and acts on potential content ciruclating online resulting from a real-world terrorism or violent extremist event. Aimed towards preventing the online proliferation of livestreamed atrocities, the protocol was created in response to the tragedy in Christchurch.

Platform invokes Content Incident Protocol

Under particular, platform-specific circumstances, a platform may alert the relevant authorities of criminal behaviour.

This is a sensitive area, as platforms are keen not to become subject to legal liability in the event that they fail to detect or pass on details of criminal activity.

Platform alerts law enforcement

Platform alerts safety partners

Where new child sexual abuse material or terrorist content emerges, platforms that are members of the NCMEC and GIFCT consortia will contribute hashes to the shared databases.

This contribution is entirely voluntary, and dependent on the platform’s proprietary definitions of terrorism and child sexual abuse. Platforms are under no legal obligation to remove content when a match is found.

Content added to known bad database

This data is retained in order to train a platform’s AI/MI tools for harm detection.

Architectural overlay

Ofcom

Phase 4

Respond

The ‘Respond’ model provides detail on how platforms respond to the finding of harmful content. It covers the process of content removal, the application of user sanctions and the procedures in place to ensure that the relevant authorities are alerted.

Key

Step in user journey

Platform activities

- security
- machine learning
- database/ persistence
- cloud
- external service
- search/analysis
- harmful content
- profile/registration
- image recognition
- management information capture

User creates or shares harmful content

User notified of content removal

Sometimes user that originated the report is notified of the action a platform takes in response.

Some platforms have spoken of the benefits of maintaining this feedback loop. Where a user is able to see follow through and affirmative action as a result of signalling harm to the platform, they are more likely to do so again.

User responsible for the report notified of follow-through

Some time later

Platform removes content

number of removals

Where content is not a direct violation of a platform’s policies but is borderline, it will be relegated in content recommendation algorithms.

Indicators of inauthenticity are commonly referred to as ‘co-ordinated inauthentic behaviour’ (CIB) in industry.

CIB can consist of liking too many things in too short a period, posting too much in too short a period, or having the same profile information as other accounts.

This behaviour points to the illegitimacy of the account and the interference of bots for the purposes of spam. CIB is the most common reason for account suspension.

Indication that the account is not legitimate

User violates community guidelines in a particularly egregious way

User repeatedly violates community guidelines

The definition of an egriegious violation and the threshold number of violations to require account suspension varies cross-platform.

On community-moderated platforms, the standard is defined by the user community.

Other platforms implement a ‘three-strike’ system, where a user receives two warnings for inappropriate behaviour before their account is suspended.

All platforms recognise the importance of consistent enforcement of user sanction.

The ultimate punitive measure platforms take is account deletion.

Depending on policy and their tolerance of bad behaviour howver, some platforms may implement temporary account suspensions. This would mean that the user subject to the measure would be unable to access their account for a specified period of time.

User notified of account deletion

number of account suspensions

Platform deletes user account

Account details are preserved to prevent ban evasion.

In some cases, platforms record IP addresses to prevent banned users from rejoining the platform.

Facebook operate the additional parameter of cookie detection to check on returning and duplicate users.

Content added to known bad database

This data is retained in order to train a platform’s AI/MI tools for harm detection.

The known bad database is sometimes referred to as a Content Bank. It does not necessarily contain the bad content in its original form - more often it contains digital representations of the content known as embeddings.

Platform outsources response

The Content Incident Protocol is a formalised playbook of actions formulated by the Global Internet Forum to Counter Terrorism (GIFCT).

It is the process by which a platform becomes aware of, quickly assesses and acts on potential content ciruclating online resulting from a real-world terrorism or violent extremist event. Aimed towards preventing the online proliferation of livestreamed atrocities, the protocol was created in response to the tragedy in Christchurch.

Platform invokes Content Incident Protocol

Under particular, platform-specific circumstances, a platform may alert the relevant authorities of criminal behaviour.

This is a sensitive area, as platforms are keen not to become subject to legal liability in the event that they fail to detect or pass on details of criminal activity.

Platform alerts law enforcement

Platform alerts safety partners

Where new child sexual abuse material or terrorist content emerges, platforms that are members of the NCMEC and GIFCT consortia will contribute hashes to the shared databases.

This contribution is entirely voluntary, and dependent on the platform’s proprietary definitions of terrorism and child sexual abuse. Platforms are under no legal obligation to remove content when a match is found.

Architectural overlay

Ofcom

Phase 5
Comply

The ‘Comply’ phase involves the processes the platform must undertake to fulfil its auditing and reporting responsibilities. It covers the management of the appeals process, the maintenance of a database of harmful content and activity, management information data capture and the publication of the transparency report.

Key

Step in user journey

Platform activities

User notified of content removal/ misclassification

The availability of appeal is dependent on two things.

Firstly, the type of violation. Removals relating to child sexual abuse material and terrorist content for instance would not be appealable, however removals on the basis of hate speech policies, spam or nudity may be.

Secondly, the type of content removed. As a general rule, posts, profile, page and group removal is appealable.

There are some exceptions to this, however. On one particular platform involved in our investigation for example, users are given the opportunity to appeal the deactivation of an account, but are unable to appeal the removal of individual pieces of content.

User is presented with ability to appeal moderator's decision

The user must act within a specified time frame.

User takes no action

User appeals moderator's decision

User notified of the reinstatement of the content

User notified that the content will retain orginal classification

Several platforms have been leaning more heavily on automated content assessment tools during the COVID-19 pandemic, due to the need to send in-house content moderators home.

This has temporarily increased the quantity of appeals and reinstatements.

Platform reinstates content

Human review

Appeals are usually sent directly to human review. Automated tools are far from matching or surpassing human capabilities in the assessment of complex, contextual cases.

A more detailed, descriptive version of the publicly available community guidelines is provided to human reviewers to assist in their decision.

This document serves as a comprehensive collection of the precedents set by previous decisions and a record of evolving policies.

Decision overruled, content has been misclassified

Original decision supported

No further action

Private information can be released following appropriate legal process such as a subpoena, court order, other valid legal process, or in response to a valid emergency request.

Law enforcement requests for information

User activity, removed content and user appeal histories are all stored by the platform. Deleted content is preserved typically for 30-90 days.

This information serves as an input into the content moderation process (to lend credibility or doubt to reports of wrongdoing) and is preserved as training material for the AI algorithms.

For more information on user data retention periods, please refer to the report.

Maintain database of user activity

Management information is gathered by platforms throughout the content moderation regime. This gives insight into the efficacy of their processes. Metrics typically gathered include:

- Number of pieces of content removed
- Detection source of removed content
- Views of removed content
- Removal reason
- Country/Region of removed videos
- Number of appeals
- Number of reinstatements

One particular platform is leading in the creation and publication of prevalence statistics. This is the most accurate measure of the efficacy of content moderation processes, as it uses randomised sampling to account for harmful content that is not currently being addressed through existing measures.

Audit performance

Most platforms publish transparency reports. These publicly available resources provide insight into the enforcement of a platform's community guidelines, requests for user data and takedown requests from law enforcement and third parties.

Platforms are under no legal obligation to publish a transparency report, but it is a widely accepted industry norm that fosters trust in online platforms.

Transparency reports are published annually on a company's website.

Publish transparency report

Architectural overlay

Ofcom

Phase 5
Comply

The ‘Comply’ phase involves the processes the platform must undertake to fulfil its auditing and reporting responsibilities. It covers the management of the appeals process, the maintenance of a database of harmful content and activity, management information data capture and the publication of the transparency report.

Key

Step in user journey

Platform activities

- security
- machine learning
- database/ persistence
- cloud
- external service
- search/analysis
- harmful content
- profile/registration
- image recognition
- management information capture

User notified of content removal/ misclassification

User is presented with ability to appeal moderator's decision

The availability of appeal is dependent on two things.

Firstly, the type of violation. Removals relating to child sexual abuse material and terrorist content for instance would not be appealable, however removals on the basis of hate speech policies, spam or nudity may be.

Secondly, the type of content removed. As a general rule, posts, profile, page and group removal is appealable.

There are some exceptions to this, however. On one particular platform involved in our investigation for example, users are given the opportunity to appeal the deactivation of an account, but are unable to appeal the removal of individual pieces of content.

The user must act within a specified time frame.

User takes no action

User appeals moderator's decision

number of appeals

User notified of the reinstatement of the content

User notified that the content will retain orginal classification

number of reinstatements

Platform reinstates content

Several platforms have been leaning more heavily on automated content assessment tools during the COVID-19 pandemic, due to the need to send in-house content moderators home.

This has temporarily increased the quantity of appeals and reinstatements.

Human review

Appeals are usually sent directly to human review. Automated tools are far from matching or surpassing human capabilities in the assessment of complex, contextual cases.

A more detailed, descriptive version of the publicly available community guidelines is provided to human reviewers to assist in their decision.

This document serves as a comprehensive collection of the precedents set by previous decisions and a record of evolving policies.

Decision overruled, content has been misclassified

Original decision supported

No further action

Private information can be released following appropriate legal process such as a subpoena, court order, other valid legal process, or in response to a valid emergency request.

Law enforcement requests for information

number of legal requests for information

User activity, removed content and user appeal histories are all stored by the platform. Deleted content is preserved typically for 30-90 days.

This information serves as an input into the content moderation process (to lend credibility or doubt to reports of wrongdoing) and is preserved as training material for the AI algorithms.

For more information on user data retention periods, please refer to the report.

Maintain database of user activity

Management information is gathered by platforms throughout the content moderation regime. This gives insight into the efficacy of their processes. Metrics typically gathered include:

- Number of pieces of content removed
- Detection source of removed content
- Views of removed content
- Removal reason
- Country/Region of removed videos
- Number of appeals
- Number of reinstatements

One particular platform is leading in the creation and publication of prevalence statistics. This is the most accurate measure of the efficacy of content moderation processes, as it uses randomised sampling to account for harmful content that is not currently being addressed through existing measures.

Audit performance

Most platforms publish transparency reports. These publicly available resources provide insight into the enforcement of a platform's community guidelines, requests for user data and takedown requests from law enforcement and third parties.

Platforms are under no legal obligation to publish a transparency report, but it is a widely accepted industry norm that fosters trust in online platforms.

Transparency reports are published annually on a company's website.

Publish transparency report

Architectural overlay