Digital Trust
& Safety Partnership

Trust & Safety Glossary of Terms

Public Consultation

January 2023

Table of Contents

Introduction

As the Trust & Safety discipline grows — in significance, complexity, and number of practitioners — there is a corresponding need to ensure a common understanding of key terms used by the people who work to keep digital services safe. Although companies have used combinations of people, processes, and technology to address content- and conduct-related risks for years, this field, following the trajectory of other technology specialty areas like cybersecurity and privacy, has reached a critical point where it has begun to formalize, mature, and achieve self-awareness.

Important discussions are happening all around the world, in homes, schools, and businesses, and at all levels of government, about what Trust & Safety should look like to best serve society and its particular relationship to the internet. But meaningful discussion is difficult without a shared vocabulary, which has been lacking.

Over the past year, the Digital Trust & Safety Partnership (DTSP) has been working to develop the first industry glossary of Trust & Safety terms, which we are publishing for consultation. Led by DTSP co-founder Alex Feerst, this glossary has the following objectives:

1. Aid the professionalization of the field and support nascent Trust & Safety teams as they build out their operations;
2. Support the codification of agreed interpretations of critical terms used across the industry; and
3. Facilitate informed dialogue between industry, policymakers, regulators, and the wider public.

The goal for this first consultation draft has been to describe how key terms are used by practitioners in industry. These are not legal definitions, and their publication does not imply that every DTSP partner company agrees with every term. We look forward to input from all interested parties to help improve this resource, which we will iterate upon going forward.

Feedback should be sent to consultation@dtspartnership.org by March 15, 2023.

I) Content Concepts and Policies

A

Acceptable Use Policy

The set of conditions and limitations governing use of a digital service that an end user or business customer (who may also agree to pass such obligations downstream to end users) must agree to as a condition of use. These are generally written in plain and concrete language (compared to legal language used in terms of service) and may also be called rules, community guidelines, or content policies).

Active Registered Users

A metric for assessing the volume of use of an online service — the number of unique users with a registered account who logged into (or otherwise authenticated a visit) and engaged with a given website, mobile website, desktop or mobile application, within a given time window, often monthly, though some services may use daily measurements for increased precision and deduplication.

Advertisement

Commercial content carried by a platform in exchange for payment. Advertisements may or may not be targeted to certain users or communities based on a variety of factors, including the demographics of a given user (either collected or inferred), or via placement next to relevant topical content.

Anonymity

A user account where the user's real identity is unknown or not displayed. This can apply in relation to one or multiple other actors, such as an online service where a user's real identity is unknown to other users and non-user third parties, but known by the service provider and not displayed, in contrast to a user whose identity is also unknown to the service provider. Some services do not allow anonymity and follow a real-name policy under which a user's name must match an official government identification.

Appeal

Loosely modeled on legal process, an appeal occurs when a user or reporter of content or conduct affected by a company's choice to disable or restrict (or decline to disable or restrict) access to an account or product, service or feature, or to take down, block, hide or classify content (or decline to do any of these), challenges that decision, and requests review, usually by another party not involved in the original decision, such as a manager or escalations department, or in some cases a designated outside body.

C

Community Guidelines

The set of conditions and limitations governing use of a digital service that a user must agree to as a condition of use. These are generally written in plain and concrete language (compared to legal language used in terms of service). Also called "acceptable use policy," or content policies).

Community Moderation

A method of content moderation whereby the users of a site or service (as opposed to site administrators or corporate employees or contractors) play a substantial role in reviewing and taking moderation actions on user-generated content. Community moderation may be a method of enforcing general sitewide community guidelines, or more specific rules or guidelines particular to a subpart of a service that the users have written independently. Community moderation has its origins in early-internet message board culture and is one of the oldest forms of online content moderation.

Content Moderation

The act of reviewing user-generated content to detect, identify or address reports of content or conduct that may violate applicable laws or a digital service's content policies or terms of service. Content moderation systems often rely on some combination of humans and machines to review content or other online activity with automation executing simpler tasks at scale and humans focusing on issues requiring attention to nuance and context. The remedies resulting from violation of a service's policy can include disabling access to content, temporary or permanent account suspension, and demotion of distribution in search or recommendation engines, and other safety interventions such as those identified in Section III below.

Copyright

A type of intellectual property law that gives the owner of literary, artistic, musical, and other kinds of expressive work certain rights to authorize how that work can be used or copied. Use of a copyrighted work without the rightsholder's permission (or without some other applicable legal basis) may constitute infringement, which can give rise to civil or criminal liability. Online services can be misused by users to infringe others' copyrights, so many services work to discourage and mitigate such abuse. Depending on factual context and jurisdiction, an online service may also be legally liable for copyright infringement facilitated by that service.

E

Engagement

Online interactions among users, content, and other elements of an online service. The defined elements of online engagement depend on the particular service and include a range of actions such as clicks, likes, comments, follows, and other forms of responsive action.

Explicit Content

Online content describing or depicting things of an intimate nature. Depending on cultural context, this may include nudity, parts of the body not generally exposed in public, sexually explicit material, or depictions of sex acts. Sometimes used interchangeably with "adult," "intimate" or "NSFW" ("Not Safe for Work"), and may also include offensive, graphic, or violent content, or association with content or commerce involving gambling, sex, cosmetic procedures, recreational drug use.

I

Identity

A user's identity online reflects their online persona including the tone and character of the content that the user shares in online networks, but it also encompasses the unique data points that are used to authenticate that specific individual within a computer network. For example, biometric data, email accounts, passwords, and two-factor authentication are all used to confirm an individual's identity when using an online service.

M

Meme

Based on the broader concept of a "meme" as a single unit of culture, thought, or behavior, internet memes are shared units of culture in the form of image, text snippet, or video excerpt generally intended to be humorous.

Message Board

An online discussion forum, usually organized around a specific topic or sub-topics, that allows users to post relevant contributions. Contributions are typically asynchronous rather than real-time, and organized into "threaded" or "nested" conversations. Message boards frequently employ community moderation to keep entries organized and on-topic.

Monetization Strategy

The way a particular product or service attempts to earn revenue, such as advertising, subscription, purchases of upgrades or virtual goods, or other methods.

O

Online Marketplace

An online service that connects merchants with consumers and generally integrates payment capability in order to facilitate the sale or purchase of goods and/or services.

P

Personally Identifiable Information (PII)

Information that can be used to identify or locate a specific individual. Personally identifiable information includes information that can directly identify a person, such as a government identification number, or information that can be used in combination with other data points, such as location data from a mobile phone used in conjunction with an individual's home and office location. As a result, a given piece of information may or may not constitute PII depending on context and what other data is associated with it.

Platform Health

The condition of an online platform's competence to adjust in response to abuse, including attacks involving fraud, disinformation, or spam (sometimes called platform integrity, depending on the service).

Pseudonymity

The use of a fictitious name, different from one's legal name, in order to, for example, conceal one's identity or protect one's privacy. Pseudonymity is different from anonymity because, through consistent use, a pseudonymous author can build a reputation around their identity and the actions that they perform under it.

R

Real-Name Policy

Requirement by a digital service that all users must self-identify using their legal name. Where platforms have such a policy, they may issue account challenges to suspect accounts which might require users to upload official photo identification matching their account name.

Right to be Forgotten

The right to be forgotten refers to the right of individuals to request that online services remove certain content related to them. For example, this may include the request to erase an individual's personal data and may apply where a search engine returns information that is inaccurate or irrelevant, and the publication of such information is not in the greater public interest. The concept is rooted in the concern that a single event, memorialized online, may impose unduly punitive consequences for a person's reputation, indefinitely, if there is no mechanism for reconsideration and removal. The right was first recognized in a 2014 ruling by the European Court of Justice, and was later codified as a "right to erasure" with the passing of the General Data Protection Regulation (GDPR) in 2018. A right to erasure has since been recognized in other jurisdictions, including Argentina, Russia, and the Philippines.

Risk Assessment

An analysis (similar to a threat model) that evaluates the types, potential severity, and likelihood of harms which may be associated with a given product, service or feature. A risk assessment may evaluate exposure to economic, legal, or brand damage; it may also evaluate potential harms to persons, as in the case of data breaches. Risk assessments are often paired with treatment plans to mitigate, eliminate or respond to unacceptable risks.

S

Safe Harbor

Generally in the law, the ability to avoid liability when certain conditions are met. In the online services context, key safe harbors under U.S. law include the Digital Millennium Copyright Act (DMCA) safe harbor (17 U.S.C. § 512) afforded to intermediaries accused

of copyright infringement for third-party content, provided they comply with the required provisions, and the broad immunity provided for interactive computer services under Section 230 of the Telecommunications Act (47 U.S.C. § 230).

Search Engine Optimization

The use of techniques that aim to increase the ranking of content in a search engine's results page and therefore be more likely to be seen by more people searching for a given term.

Social Media

A service that forms an online network enabling individual users to interact with other users with common features such as posting text, images, and video, liking or commenting on others' activity, or otherwise engaging with others online. In contrast to a messaging application, social media tends to default to public, such that any other user (or potentially anyone online) can interact with others, depending on the product and privacy features of a particular service.

T

Terms of Service

The legal agreement between a user and a service provider under which the user uses the service. A service's rules and policies (such as community guidelines, acceptable use policy, content policy, or other rule sets) are often set forth in a separate document but incorporated by reference into this legal agreement, making compliance with such rules part of using the service.

Trademark

A type of intellectual property law that protects any word, symbol, or expression that uniquely identifies a product, service, or company. The mark helps consumers know that they are purchasing the authentic product or service offered by the trademark holder. Impersonation of brand accounts (whether parody or not) often implicates trademark law, as it may create confusion over whether a given statement or other content originated from or is endorsed by a brand.

Transparency Report

A report periodically issued by a service that discloses metrics and insights about its approach to salient risks and relevant enforcement practices, including how it has handled requests to remove or restrict user content, and requests for user data. Transparency reports often detail government requests for user records, providing greater public transparency around which kinds of private information governments have requested, under what authority, and how frequently; the reports may also disclose how much content was removed due to various legal provisions such as copyright, including fraudulent takedowns, and other forms of abuse.

Trust & Safety

The field and practices employed by digital services to manage content- and conduct-related risks to users and others, mitigate online or other forms of technology-facilitated abuse, advocate for user rights, and protect brand safety. In practice, Trust & Safety work is typically composed of a variety of cross-disciplinary elements including defining policies, content moderation, rules enforcement and appeals, incident investigations, law enforcement responses, community management, and product support. Since about 2005, it has developed into a distinct profession in its own right, with several professional organizations (such as DTSP and the Trust & Safety Professional Association) focusing on Trust & Safety functions emerging since 2020.

U

User Controls

Technical measures designed to allow users to control their own product experience where possible and appropriate, such as blocking or muting other users or certain types of content, expressing preferences for use of private information, and adjusting security settings. Also called "user settings."

V

Verification

The process by which a company confirms the identity or authority of a user or other pertinent facts associated with a user account, which may be controlled by a range of actors, including individuals, a company, an advertiser, a politician or political party, or a government agency. Verification procedures may entail uploading a government identification document or official mail (such as a utility bill) received at a specific location to confirm geographic address to the online service or a third-party provider.

Verified User

A user whose identity has been verified by the provider or a third-party, such as by examining government ID. Social media services often have public-facing verification labels to signal that the accounts of public figures, commercial brands, politicians, government actors, or celebrities are who they say they are, to encourage reliance on them as trustworthy, and prevent third-parties from using impersonation accounts in fraud or misinformation campaigns.

II) Common Types of Abuse

A

Abuse

Use of a product or service in a way that violates the provider's terms of service, community guidelines, or other rules, generally because it creates or increases the risk of harm to a person or group or tends to undermine the purpose, function or quality of the service. May also refer to using the product or service to abuse another person or a group or in ways that violate local law.

Account Takeover

The scenario where an unauthorized user gains control of a user account, through means such as hacking, phishing or buying leaked credentials.

Astroturfing

Organized activity intended to create the deceptive appearance of broad, authentic grassroots support or opposition to a given cause or organization, when in reality the activity is being motivated, funded or coordinated by a single or small number of obscured sources.

B

Brigading

Coordinated mass online activity to affect a piece of content, or an account, or an entire community or message board, for example by upvoting or downvoting a post to affect its distribution, mass-reporting an account (usually falsely) for abuse in an attempt to cause the service provider to suspend it, or inundating a business with good or bad reviews.

C

Catfishing

The scenario where someone creates a fake persona on an online service, such as social media or a dating application, and forms a relationship with someone who believes the persona to be real. This behavior is often associated with financial fraud and other forms of exploitation of the victim.

Content- and Conduct-Related Risk(s)

The possibility of certain illegal, dangerous, or otherwise harmful content or behavior, including risks to human rights, which are prohibited by relevant policies and terms of service.

Coordinated Inauthentic Behavior

Organized online activity where an account or groups of accounts including "fake" secondary accounts (which exist solely or mainly to engage in such campaigns) act to mislead people or fraudulently elevate the popularity or visibility of content or accounts, such as mass-following an account to raise its clout. In some cases, a single, hidden source or organization will deploy many fake accounts in order to create a false appearance of authentic and credible activity. In other cases, people using their own, real accounts will coordinate online to achieve a misleading purpose, such as the appearance that a view of belief is more widespread than it is, or to cause wide distribution of a particular piece or type of content. Sometimes called "platform manipulation" or "content manipulation."

Copyright Infringement

The use of material that is protected by copyright law (such as text, image, or video) in a way that violates the rights of the copyright holder, without the rightsholder's permission and without an applicable copyright exception or limitation. This can include infringing creation of copies, distribution, display, or performance of a covered work, or the unauthorized creation of derivative works. Infringement may involve primary liability (for the person who did the infringing conduct) or secondary liability for others involved in that conduct (such as a hosting company whose service hosts images posted by a user). A digital service hosting user-generated content receives safe harbor under Section 512 of the Copyright Act (17 U.S.C. § 512), so long as it complies with the applicable notice and takedown procedures set forth in that law.

Counterfeit

The unauthorized manufacture or sale of merchandise or services with an inauthentic trademark, which may have the effect of deceiving consumers (or people observing consumers) into believing they are authentic. The manufacture or sale of counterfeit goods is a form of trademark infringement and secondary liability for this conduct is a concern for online marketplaces.

Cross-Platform Abuse

Instances where a bad actor or group will organize a campaign of abuse (such as harassment, trolling or disinformation) using multiple online services. This has the effect of making it more difficult and time-consuming for affected persons to have the abusive content removed, as they will be required to contact each service separately and explain the situation. Sometimes, the same content will simply be re-posted across multiple platforms. In other cases, bad actors will divide content or conduct such that no one service carries the full abusive content. As a result, lacking full context of the entire campaign, or if a service's policy restricts its inquiry only to content or conduct that directly involves that service, a given service may determine that no violation has taken place. Typically such situations require research and integration of data from multiple services, and investigation of the background context of the bad actor(s) and affected person(s) to make more meaningful assessments and respond appropriately.

Child Sexual Abuse Material (CSAM)

Sexual content produced by exploiting underage subjects, with related subcategories including CAI ("Child Abuse Images"), CEI ("Child Exploitative Imagery"), CAM ("Child Abuse Material"), CSEM ("Child Sexual Exploitation Material"), CSEAI ("Child Sexual Exploitation and Abuse Imagery"), PFP ("Perceived First Person") CSAM, SG ("Self-Generated") CSAM, and "Child Pornography." "Simulated Child Pornography" contains modified or invented depictions of children without the direct involvement of any underage subjects. The industry discourages the use of the term "Child Pornography," which is still used as a legal term, including in the United States. CSAM is illegal in nearly all jurisdictions, making detection and removal of CSAM a high priority for online services.

D

Defamation

A legal claim based on a false statement asserting a fact about a person that is shared with others and which causes harm to the reputation of the statement's subject (the legal elements and applicable defenses vary by jurisdiction). Defamation can be conveyed through a range of media, including visually, orally, pictorially or by text. In the United States, supported by First Amendment jurisprudence, the burden of proof to establish defamation is on the person alleging they have been defamed. In other jurisdictions, such as Europe, the burden of proof may be on the defendant to establish they did not commit defamation. These differences in legal approach and levels of associated legal risk may influence the takedown processes for defamation disputes adopted by online services in various localities.

Dehumanization

Describing people in ways that deny or diminish their humanity, such as comparing a given group to insects, animals or diseases. Some experts in this area cite dehumanizing speech as a possible precursor to violence (sometimes to the point of genocide), because it may make violent action seem appropriate or justified against "non-human" or "less-than-human" targets.

Denial of Service (DOS) / Distributed Denial of Service (DDOS)

A malicious attempt to disable an internet server, service, or network by sending it a large volume of traffic to overwhelm its capacity. In the course of the attack, the attacker may infect other networks with malware in order to hijack control of them; these hijacked bot networks may then be used to attack the target by sending internet requests to the target's IP address.

Disinformation

False information that is spread intentionally and maliciously to create confusion, encourage distrust, and potentially undermine political and social institutions.

Doxxing

The act of disclosing someone's personal, non-public information — such as a real name, home address, phone number or any other data that could be used to identify the individual — in an online forum or other public place without the person's consent. Doxxing may lead to real world threats against the person whose information has been exposed, and for this reason it is often considered a form of online harassment. Some services may also consider aggregating and disclosing publicly available information about a person in a menacing manner sufficient to constitute doxxing.

F

Farming

Content farming involves creating online content for the sole or primary purpose of attracting page views and increasing advertising revenue, rather than out of a desire to express or communicate any particular message. Content farms often create web content based on popular user search queries (a practice known as "search engine optimization") in order to rank more highly in search engine results. The resulting "cultivated" content is generally low quality or spammy, but can still be profitable because of the strategic use of specific keywords to manipulate search engine algorithms and lead users to navigate to a page, allowing the owner to "harvest clicks" for ad revenue. Account farming involves creating and initially using accounts on services in apparently innocuous ways in order to build followers, age the account, and create a record making the account appear authentic and credible, before later redirecting the account to post spam, disinformation, or other abusive content or selling it to those who intend to do so.

G

Glorification of Violence

Statements or images that celebrate past or hypothetical future acts of violence. Some online services restrict or prohibit glorification of violence (including terrorism) on the reasoning that it may incite or intensify future acts of violence and foster a menacing or unsafe online environment, though it is challenging to distinguish glorification of a subject from other types of discussion of it.

H

Harassment

Unsolicited abusive behavior against another person, often repetitive and usually with the intent to intimidate or cause emotional distress. Online harassment may occur over many mediums (including email, social media, and other online services) and it may expand to include real world abuse. Online harassment may take the form of one abuser targeting a person or group with sustained negative contact, or it may take the form of many distinct individuals targeting an individual or group.

Hate Speech

Abusive, hateful, or threatening speech that expresses prejudice against a group or a person due to membership in a group, which may be based on legally protected characteristics, such as ethnicity, sex or gender identification or sexual orientation.

I

Impersonation

Online impersonation most often involves the creation of an account profile that uses someone else's name, image, likeness or other characteristics without that person's permission to create a false or misleading impression that the account is controlled by them.

Incitement

To encourage violence or violent sentiment against a person or group.

M

Misinformation

False information that is spread unintentionally and usually not maliciously, which may nonetheless mislead or increase likelihood of harm to persons. (Compare with "disinformation.")

N

Non-Consensual Intimate Imagery

Non-consensual image sharing, or non-consensual intimate image sharing (also called "non-consensual explicit imagery" (NCEI) or "revenge porn"), refers to the act of creating, publishing or sharing an explicit image or video without the consent of the individuals visible in it. Non-consensual intimate imagery (NCII) may contain nudity or sexually explicit acts. The imagery may have been created by or with the consent of the individuals shown, such as in the context of an intimate relationship, or created without consent through the use of hidden cameras or other surveillance methods. Similarly, it may have been obtained with or without consent to possess it, or consent to possess it may have been revoked. Sharing or distributing the content to others, without the depicted person's consent, is widely regarded as a form of harassment and is illegal in some jurisdictions. It should be noted that non-consensual intimate imagery is distinct from the unlicensed sharing of copyrighted, commercially-produced professional adult content.

S

Sock Puppets

Multiple, fake accounts used to create an illusion of consensus or popularity, such as by liking or reposting content in order to amplify it.

Spam

Unsolicited, low-quality communications, often (but not necessarily) high-volume commercial solicitations, sent through a range of electronic media, including email, messaging, and social media. It dates back to 1980s BBS (Bulletin Board System, an early form of online communities) slang for repeated posting in volume to flood out rival messages, derived in turn from a Monty Python comedy sketch in which "spam" is said over a hundred times.

Synthetic Media

Content which has been generated or manipulated via algorithmic processes (such as artificial intelligence or machine learning) to appear as though based on reality, when it is in fact artificial. Generally, synthetic or manipulated media (including "deepfakes"), may be used within the context of abuse to deceive or cause harm to persons, such as causing them to appear to say things they never said, or perform actions which they have not (as in the case of "deepfake pornography").

T

Troll

A user who intentionally provokes hostility or confusion online.

Terrorist and Violent Extremist Content (TVEC)

A type of content that is increasingly a focus of lawmakers and regulators concerned with preventing its availability. Approaches to defining the category vary, including actor- and behavior-based frameworks, and in order to detect and remove it, online services may rely on research and lists of terrorist or extremist organizations created by subject matter expert organizations, such as the United Nations Security Council's sanctions list.

V

Violent Threat

A statement or other communication that expresses an intent to inflict physical harm on a person or a group of people. Violent threats may be direct, such as threats to kill or maim another person; they may also be indirectly implied through metaphor, analogy or other rhetoric that allows the speaker plausible deniability about their meaning or intent. Often overlaps with incitement, such as making a public statement that a person deserves to be harmed, implicitly encouraging others to do so.

III) Enforcement Practices

A

Account Suspension

A penalty imposed on a user's account that temporarily restricts part or all of the service's functionality, often due to a policy or terms of service violation. An account suspension is generally temporary, though in egregious cases may be permanent. Reactivation may be dependent on the user completing some action, such as making a payment, passing authenticity verification, or removing or modifying content. Or, reactivation may occur after a set amount of time without additional action.

Account Termination

Closing a user account, generally permanently. An account termination may be requested by a user who no longer wishes to maintain an active account on the online service. Or, it may be initiated by the provider after severe or repeated violations of the service's acceptable use policy or Terms of Service. Account termination will usually trigger deletion of the user's account data after the period allotted for appeals or reconsideration has passed. Certain privacy laws, such as the EU's General Data Protection Regulation (GDPR), provide users with a right to request deletion of personal information associated with them. In some cases, the data associated with a terminated account may be preserved, such as when the provider has received a notice to preserve account data from law enforcement or a potential civil litigant.

Age Assurance

A service unwilling to rely solely on users' self-reported age for compliance, due to concerns about misrepresentations, may adopt more rigorous means to ascertain an account holder's age such as requiring the user to upload an official document to prove their birth year, or using third-party data to cross check against public records.

Age Verification

A process in which a visitor to a website or user of an online service is asked to verify that they meet a particular minimum age requirement. Age verification may rely on self-reported data, such as a simple self-reporting interstitial that asks the user to self-affirm their age by checking a box that they meet the appropriate age requirement to proceed, or rely on the provision of government identification to the service provider or a third party (also known as "age assurance"). In many jurisdictions, age verification is implemented to protect minors from online content or services that lawmakers have deemed inappropriate for their maturity level, such as internet gambling, pornography, or alcohol sales. It is also important for compliance with online privacy laws that regulate the collection and processing of data from minors.

Anti-Abuse

Anti-abuse teams work to prevent, detect, and mitigate uses of digital products and services that may be unauthorized, harmful or illegal under local or international laws. May also be known as safety teams, or various similar monikers.

B

Ban

A decision made by a service to disallow or restrict access to a type of content, or to prevent a user account or group from using a service, such as account suspension or termination. Related slang includes referring to a permanent ban of a user as a "perma-ban" or "deplatforming" and referring to a moderator's power to ban a user as the "ban hammer."

Block

The disruption of an online service or particular parts of a service. As a feature, a user can have the ability to "block" the exposure of their content to other user(s) on the platform, or conversely, eliminate or reduce their own contact with certain content or users. Reasons for this may include a past history of harassment or simply because a user doesn't want to see a particular type of content. Depending on the nature of the service and the exact scope of features included as part of the user blocking function, a block may disable other features, such as preventing the blocked user from seeing the blocker's account profile or other associated information. Blocking may also refer to the scenario where a government or internet service provider disables access to certain services within its borders or service area, respectively.

C

Content Removal

The process by which content is removed or deleted from an online service. Content removal may be done by the user who posted the content or by the online service hosting it. When content is removed by the online platform for violating applicable law or company policy, the user generally receives a notice of the removal which may include the reason for removal. There may be an appeals process by which a user may request an explanation for the removal and petition for review of the content in order to reinstate it.

D

Deindexing

Removal of content from a search engine or other directory so that it does not appear in results, generally making it harder to find and less likely to be viewed as a result. Deindexing may occur with or without removal of the content itself from the digital service where it is hosted.

Demonetization

Removal of the ability for a piece of content or a user account to earn advertising or other revenue on a platform. Demonetization may be a remedy for misconduct or result from a piece of content being deemed insufficiently abusive to remove, but inappropriate to encourage through monetization. Content that is within platform terms, but deemed not to be brand-safe for advertising may also be demonetized.

Downranking

In the context of search results, or distribution of content within social media feeds, downranking usually refers to manual or automated reduction in visibility or relevancy of a piece of content or user account, such that it either doesn't appear or appears lower down in results than other items. Downranking may be used in some cases as a tool to limit the visibility and spread of abusive content.

E

Escalation

In the Trust & Safety context, an organizational process in which an alleged violation of an online service's policies or the terms of service is sent to a higher managerial level for additional review. Generally, when a case is escalated it goes to a more experienced or authoritative person or team to conduct further review and analysis.

F

Fact-Checking

A process of factual verification applied to published statements in order to provide an accurate, unbiased analysis of whether the claims in the communication can be confirmed, and thus trusted. In the online context, some social media platforms hire qualified fact-checking individuals or organizations to conduct analysis of claims that have been posted online in order to limit the spread of misinformation. In some cases fact-checking organizations will rate a communication as "true" or "false;" in other cases the fact-checking system will offer additional ratings, such as "needs more context."

Filtering

A usually automated process where content that matches given parameters may be automatically removed, downranked, or have distribution disabled. Filtering may be based on exact or "fuzzy" keyword matching, pattern matching against regular expressions, or hashes that uniquely identify a given piece of content. In addition to evaluating the content itself, filtering may also make use of behavioral signals associated with user accounts. Scoring systems may be applied to undesirable behaviors (such as those associated with known spam accounts), and thresholds set by system administrators above or below which filtering actions may be initiated. Filtering may be performed by a range of actors including private companies, such as digital services or internet service providers, or by government actors.

Flagging

Also known as reporting, the process by which a user, Trust & Safety agent, algorithm, or partner organization may request review of online content, conduct or a user account. In some cases, and depending on the service and other context, flagging may automatically trigger the temporary suspension of account services or benefits while the issue is being investigated.

Freezing

An administrative action taken by a digital service to prevent a user from using or taking further action with their account, as in to "freeze an account." Also called "locking."

G

Geoblocking

A technological feature that restricts access to online content or a service, based on the user's geographic location. Geoblocking (or geofencing) may be used when a government has prohibited access to certain content or activities from its territory (such as sexually explicit content or online gambling), or a particular piece of content is restricted by a private licensing agreement (such as media streaming rights, which may differ from place to place). Geoblocking is typically used when the content otherwise does not violate a platform's terms, and is used to keep the content available to users in jurisdictions where it is not prohibited. As geofencing is often based on user IP address, use of a VPN may allow a user to spoof their location and access content or services that are geoblocked in their country.

I

Interstitial

An "in-between" page or message that appears before a user's targeted destination, acting as an interruption to present or obtain information for some purpose, such as to conduct an age verification, display an advertisement, present a content warning, or show the user another type of notification before they are relayed onward to the destination.

M

Mute

A softer alternative to blocking, in the context of social media, muting generally refers to user-controlled tools which allow users to stop receiving distributions of content published by selected accounts, or by keywords. Unlike blocking, where a blocked account may not be able to see the activity of the person who blocked them, an account subject to muting by another generally will not have their own user experience otherwise affected or know they have been muted.

N

Notice to Users

Many services as a matter of policy provide notice to users whose content or conduct is at issue in an enforcement action (with relevant exceptions, such as where prohibited by law or in order to prevent harm to a person) and where relevant to the complainant who reported it. Another form of notice is in-product indicators of enforcement actions taken, including broad public notice (e.g., a "tombstone" screen presenting relevant information where content has been removed).

P

Parental Controls

Technical features available on some electronic devices or internet services that allow parents or others to control what online content that an associated child's account can access, or actions that can be taken, such as downloading an app or making in-app purchases.

R

Read-Only

Removal of a user account's publishing permissions, either temporarily or permanently. The account holder is typically able to still access their account for a service, and view content, but not publish content, or in some cases otherwise interact with other users.

Reinstatement

Following a suspension of content or an account from a digital service, the restoration of an account or content back to its normal state. This may result from a successful appeal or the end of a set suspension period.

S

Strikes

A Trust & Safety penalty process by which the accumulation of repeated violations of applicable law, terms of service or acceptable use policy will trigger more severe penalties against a user's account. The reasoning behind a multiple strikes system is that a user who repeatedly violates a service's rules is likely acting intentionally, and due to this malicious intent penalties imposed on the account should be enhanced to prevent likely future abuse. Relatedly, some laws encourage this type of penalty structure, such as the DMCA, which requires that online services have a "repeat infringer" policy.

T

Throttling

A method of reducing potentially negative impacts of actions taken by internet users, such that they can only perform a given number of actions (such as posting content, creating accounts, or engaging with others) within a given period of time. Once a limit is hit by a user account, that user is not able to again perform the action until the time limit specified by system administrators has passed. Throttling is often used as one of several tools to combat unwanted automation such as spam and bot accounts. Also known as "rate-limiting."

IV) Trust & Safety Technology

A

Algorithm

Broadly defined, an algorithm is a process or set of rules for solving a given problem in computing. In the context of online services where content is displayed to users, a key use of algorithms is to select, rank, and personalize content for users. They can also be used for safety purposes, to automatically detect content that may violate terms of service. Algorithmic outputs may be based on a nearly infinite range of factors and inputs, depending on the design of a product or service and the data available. Some common (but not universal) algorithmic inputs for ranking or recommendation purposes include information known or inferred about a user, such as past interactions with the service, or demographic data about the user. Likewise, information about the content itself, such as its topic, media type, or general popularity, can be used. In the context of search services, an algorithm may be used to sort, weight, include, and exclude potential responses to a search query.

Allow List

A list of pre-approved items — such as files, individual user accounts, words, IP addresses, or URLs — that are allowed to appear or operate in a network (sometimes called a "white list" although this term is now discouraged). It generally contains a list of items that are assumed to be or have been deemed safe.

Amplification

An increase in user exposure to certain online content beyond that occurring from a service's basic hosting or transmission features. Depending on a product or service's particular design, it can occur due to a range of platform features including recommendation based on human or automated curation, engagement by users, paid distribution bought from the platform, or resulting from deliberate or inauthentic interactions such as astroturfing, brigading, or automated bot campaigns.

Automated Moderation

The use of automation technology to perform at least some part of content moderation processes. Automation may be used to detect potential abuse, through methods like keyword filtering, hash matching, behavioral analysis, machine learning, and artificial intelligence. In some cases, a human would then evaluate the potential abuse in light of the applicable company policy and determine what appropriate action, if any, to take. Automated moderation seeks to augment the effectiveness of human teams through functions such as advance surfacing or prioritizing potential problems, sorting issues into categories, suggesting a response, or in some cases, selecting and applying a rule and triggering an enforcement action.

Automation

Technologies and processes which perform online actions in order to achieve a specific goal. Automation may be used for legitimate ends (as when using an API), or for abuse such as by bot accounts, spammers, and astroturfing. It may be particularly useful in helping to scale and speed up decisions that are binary in nature, whereas more nuanced decision-making may require at least some human involvement. Automation can be deployed along a spectrum from full automation, requiring no human intervention when operating within parameters, to partial automation, which may involve as part of the system's feedback loop.

B

Bot / Bot Account

A software application that runs automated tasks over the internet. Typically, bots perform tasks that are both simple and structurally repetitive, at a much higher rate than would be possible for a human alone. A "robot" user account could be used for publishing content or interacting with other user accounts in response to some triggering event. Depending on the rules of a given service, and the nature of a bot's activity (which could be benign, harmful or neither), a bot account may be allowed or disallowed, and may be labeled as a bot account. "Botnets" are large coordinated networks of bots, run by a "botmaster."

Browser Extension

A software application which acts as a plug-in to a web browser and augments functionality natively available in the browser.

C

CAPTCHA

"Completely Automated Public Turing test to tell Computers and Humans Apart." A challenge-response system used to differentiate human interaction with a digital service from automated or bot activity, and to reduce the likelihood and impact of inauthentic activity.

D

Deny List

A list of items — such as files, individual user accounts, words, IP addresses, or URLs — that are banned from a network or service (sometimes called a "black list" although this term is now discouraged). Inclusion in the list may result from past violations of the law or service's rules or other evidence that a given user account or file may be malicious or inherently violate the service's policies or applicable law. Also known as a block list.

Device Fingerprint

A technical measure that allows merchants and others to identify an individual internet user based on aggregating data associated with the user's browser and the device (web or mobile) used. Device fingerprinting may collect data such as the user's IP address,

geographic location and time zone, VPN, or operating system. Unlike web cookies, which can be deleted and are stored on the client side (e.g., on the user's device), device fingerprinting is generally stored in a database on the server side of the transaction and is controlled by the party who is keeping it for identification purposes. Its uses include security as well as ad tracking.

Digital Rights Management (DRM)

Software used to control and restrict use of files containing copyrighted material in order to prevent infringement, though sometimes with the side effect of making it impossible to use content under the scope of legal exceptions and limitations to copyright, such as fair use.

E

Export

The transfer of data, files, or other user information from an online service to another database or storage system.

F

Hash Filtering

Hash functions are algorithms applied to inputs of variable length to provide a fixed-length output. A given hash algorithm (such as SHA-256) always returns the same value for a given input, making it a means of uniquely identifying a piece of digital content (such as an image, video, or block of text). Content hashes therefore form the basis of hash filtering, whereby items exactly matching known hashes may be detected and acted on automatically, triggering actions such as a prohibition from being uploaded or served by a system. A limitation of conventional hashing is that minor modifications to inputs will result in new unique hashes, thereby allowing nearly identical copies to defeat filtering based on previously known hashes. Perceptual hashing is a technique of detecting content that is substantially similar to known hashes, which may be escalated to human review or automatically blocked based on predefined thresholds of similarity.

I

Internet Infrastructure

The combination of internet hardware and software that support the successful delivery of online services. Elements of internet infrastructure include routers, telephone and fiber optic cables, and domain name servers.

IP Address

An internet protocol (IP) address is a numerical or sometimes alphanumeric string that uniquely identifies a device on the internet or a computer network. An IP address may be static or dynamic, and sometimes multiple devices connected to a single router or internet service (such as a VPN) may share an IP address.

K

Keyword Matching

A method of content filtering based on matching textual strings contained in keywords, or combinations of strings. Sometimes, automated actions are then applied to content that matches keyword filters based on rules determined by system administrators, such as surfacing it for human review or marking it for suspension.

L

Logged-In / Logged-Out

An account holder who has registered for an account, and is signed into it, may be considered logged-in. Functionality available to logged-in users of a platform is usually more than what is available to logged-out users, as the history and implications of prior user actions may be drawn upon.

Login Challenge

In cases where suspicious activity has been detected on an account, a login challenge may be issued in order to determine that the user in question is human (and not a bot, for example), or is who they claim to be. Login challenges may occur, for example, when a user logs in to a service from a previously unknown device, or from an IP address in a geographic location that differs from their usual account activity. A login challenge could consist of a CAPTCHA, requiring the user to confirm non-public data linked to their account (such as a phone number) through confirmation with a separate device, or by uploading official documents.

Logs

A record of prior events that occurred within a network, software application or operating system. When trying to understand an incident or malfunction, logs can be a helpful resource because they provide a detailed record of previous activity and system performance. In the context of online safety, the Trust & Safety team may log incoming complaints, decisions, and enforcement actions for many reasons, including transparency, quality assurance, training purposes, or legal compliance.

M

Messaging Applications

Applications based on person-to-person communication via text, images, video, voice or some combination, or shared within a group generally limited to a certain number, rather than publicly discoverable and potentially unlimited in the number of viewers, and in which users generally have some expectation of privacy.

T

Ticket

In the context of customer support software often used to track complaints including Trust & Safety cases, a "ticket" is shorthand for a specific support case or incident, generally logged as an ID number or other filing convention.

Tools

Trust & Safety tools refers generally to software used by moderators for detecting potential abuse (such as keyword filters, rules engines, as well as Machine Learning or Artificial Intelligence systems), flagging content or accounts to act on, implementing remedies, generating tickets, and communicating with users and complainants.

U

Uniform Resource Locator (URL)

A URL or "uniform resource locator" is the unique web address that directs a web browser to locate an internet resource. Every web page or other internet resource online has a unique IP address, composed of a numerical and/or alphanumeric string. Because the underlying IP address can be difficult to remember, the IP address is linked to a generally more memorable URL through the DNS (Domain Name System), which translates a URL name into the appropriate number to make the request.

Username

A set of characters that uniquely identifies a user on an online service, computer network or system. A username is often used in conjunction with a password to securely authenticate the user before granting access to the multi-user and/or protected system.

V

Virtual Private Network (VPN)

A "Virtual Private Network" (VPN) is a secure connection that allows public network users to connect to a private network, usually encrypting all network traffic. This has a number of uses, such as allowing a remotely located person to work "inside" a private and more secure network environment. It also can allow a web user to disguise their IP address, and therefore conceal their location from being accurately detected or logged by the provider. A "proxy," meanwhile, is a server that acts as a relay between a client and server. It too may replace a web user's IP address with another one. Unlike VPNs, which capture all network traffic, a proxy only functions at the application layer (i.e., for the application where you set up the proxy, such as a web browser), and typically does not encrypt traffic.

Index

IV) Trust & Safety Technology

- Algorithm
- Allow List
- Amplification
- Automated Moderation
- Automation
- Bot / Bot Account
- Browser Extension
- CAPTCHA
- Deny List
- Device Fingerprint
- Digital Rights Management (DRM)
- Export
- Hash Filtering
- Internet Infrastructure
- IP Address
- Keyword Matching
- Logged-In / Logged-Out
- Login Challenge
- Logs
- Messaging Applications
- Ticket
- Tools
- Uniform Resource Locator (URL)
- Username
- Virtual Private Network (VPN)