Teleperformance

Age Determination

A Best Practice Guide for **Online Platforms** that Feature User-Generated Content

Table of Contents

Context

Platforms, including social media, gaming, and dating companies, often have user-generated content (e.g., videos, images, and other content) that depicts and/or discusses children, i.e., people under the age of 18 (or some other age based on a country's laws). Platforms typically have different policies related to content featuring adults and children, based on the maturity of the theme, or other such considerations regarding age-appropriateness.

When platforms are enforcing policies related to children, it can be challenging to accurately determine the age of a person featured in an image or video. This can have implications as to whether a piece of content should be treated as an image or video of a child for the purpose of enforcing child protection policies and safeguards.

It is of critical importance that platforms have processes, technology, policies, and governance in place to properly determine the age of people depicted in content. The most critical example of this when platforms monitor sexually explicit images/videos to determine whether the material should be considered Child Sexual Abuse Material (CSAM). Platforms must have sound ways to determine if children are being depicted in any such exploitative or abusive material; this is so that they can take the appropriate actions required to report and remove such content from their platforms and prevent such content from being uploaded in the future.

A working group composed of policy, technology, and content moderation experts across the International Centre for Missing and Exploited Children, Thorn, Teleperformance, and Aylo was formed in order to understand and document best practices in age determination toward this goal.

An Introduction to Age Determination

Age determination is the means by which platforms determine whether content features or depicts a minor or child, a person under the age of 18 (or another pre-specified age). While age verification (under the umbrella of age assurance) typically focuses on determining a user's age when deciding whether an individual should be allowed to create an account or sign up for a platform's services, age determination is focused on evaluating an existing piece of content to decide whether a child is featured (discussed or depicted) in this content, primarily through visual assessments. These assessments could be conducted either by human content moderators or by using some form of technology. Other contextual analysis may also be used to aid in age determination, either in conjunction with or instead of visual assessments.

For an understanding of best practices for age assurance, we recommend reviewing the Digital Trust and Safety Partnership's report here. This paper focuses on age determination and is meant to be complementary to existing reports on related topics under the age assurance umbrella.

The Challenge

There are a number of challenges to effectively and accurately determining the age of a person in a piece of content. From an overarching level, there is no single, universally applied industry standard or globally unified process to accurately assess the age of people featured in images/videos.

From a process perspective, each platform has its own methods, including the use of technology and human review, to make this determination. Each platform typically has its own guidelines, training, and associated documentation, which can vary widely in level of detail and in the availability of localized content for moderators or others involved in the process to leverage.

From a technology perspective, new research has found that bias is also pronounced in existing machine learning (ML)/artificial intelligence (AI) tools (discussed further below). AI technologies used by platforms may be trained on inherently biased data sets that result in bias or reduce the effectiveness of estimating the age of certain ethnicities (bias inherent in AI data sets). Equally, relying solely on AI technologies can be limiting, given the trend of "aging up" (children appearing older than their age with the assistance of clothing, makeup, etc.) of young children, particularly girls.

From a people perspective, cultural norms, including clothing, hair, and makeup, can lead to some people appearing more or less mature than their stated age, resulting in underenforcement or overenforcement of content policies dependent on age. In addition, enforcement may differ based on the moderator's race/ ethnicity due to unconscious bias; therefore, relying solely on visual age assessment (as opposed to documentary evidence) can be challenging and may similarly result in underenforcement or overenforcement in certain markets.

Section 1: Overarching approach: balancing technology and human review

Most platforms use a combination of technology and human review to conduct age determination, though this may vary depending on the size and type of platform.

From a technology perspective, AI can be useful because it enables speed, scale, and proactivity of action on age determination. However, it is important to note that research has demonstrated that biases in human perception of facial age are present and more exaggerated in current AI technology. Particularly, this research shows that "AI is even less accurate and more biased than human observers when judging a person's age — even though the overall pattern of errors and biases is similar. Thus, AI overestimated the age of smiling faces even more than human observers did. In addition, AI showed a sharper decrease in accuracy for faces of older adults compared to faces of younger age groups, for smiling compared to neutral faces, and for female compared to male faces. These results suggest that our estimates of age from faces are largely driven by particular visual cues rather than high-level preconceptions."

It is important to understand how AI works for age determination to determine its relevance and usefulness for platforms that are already using —or platforms that are considering using— this as part of their solution for age determination. ML/AI for age estimation is generally built using supervised machine learning. Supervised learning is a subset of machine learning where the model is provided with a set of labeled data (known as the "training data"). This training data is labeled according to the desired classification (e.g., an age range) or regression problem (e.g., a numerical age). The model iterates over a combination of features and labels, ultimately producing a function. This function can now be employed to make predictions on novel input data.

A significant subset of research focuses on age estimation from the face of an individual depicted in an image or video. However, age estimation technology can also be built using other aspects of the available content or other types of content (e.g., the full body of an individual depicted in an image or video, CT scans, and MRIs in the medical settings, etc.). It is worth noting that the prevalence of faces in CSAM is low due to bad actor efforts to obfuscate information that could support victim identification.

Given current effectiveness and efficiency of both human and tech solutions, automated review followed by manual verification is still likely the best approach to balance accuracy and speed at scale.

Section 2: Key Technologies for Age Determination

AI technology is essential for age determination at scale. The capabilities, benefits, and limitations will be explained for each key technology here.

ML/AI for Age Estimation

Focusing on AI for age determination using faces, there are generally two steps involved in AI solutions.

1. First, the face in the image must be detected. This can be understood as a type of object detection task, where the relevant object in this circumstance is the face of the individual. An AI face detection model that has been trained on labeled data (e.g., images of people, where the face of each person in the image has been labeled using a bounding box or key points) can be used to conduct this type of face detection.
2. The face is provided as input to the age determination model. The model then outputs a prediction using the weighted function that it learned during model training.

Here, even though the method works by processing a photograph of a person's face, the only output is a non-identifying age estimation. As noted above, AI solutions for age determination encounter challenges regarding bias. Many factors can impact the performance of an ML/AI age estimation model, including factors such as gender, ethnicity, image resolution, head positioning, facial scarring, occlusion, etc.

There are several related technologies that do not directly output an age determination but can be used to support an age determination task:

1. CSAM Classifiers

An AI solution trained to detect CSAM is a different task than age determination. As the task is to predict whether a piece of content is CSAM, the training data set will need to either include CSAM or some proxy for CSAM. An example of proxy training is that a CSAM classifier could be trained as a combination of sexual content detection (e.g., trained on adult pornography and benign content) and age determination (e.g., trained on content representing a distribution of ages across children and adults).

Products like Google's Content Safety API and Thorn's Safer provide a likelihood estimate that a piece of content is CSAM, but do not make a determination of age. Notwithstanding, these types of classifiers could

support an age determination task. For example, if a CSAM classifier predicts an image to be CSAM, that content could be entered into a reviewing queue for a content moderator, indicating in advance that the content may depict a child. It is important to note that some of the technologies used by companies (such as Google's Content Safety API) require human review for everything that is flagged if the intent is to report this to the National Center for Missing & Exploited Children (NCMEC). Law enforcement is encouraging platforms to improve their age assessment accuracy before submitting reports to CyberTips in order to reduce the volume of inaccurate reports.

2. Hashing and Matching

This technology consists of both cryptographic (exact match) and perceptual (fuzzy match) hashing solutions where an image or video is hashed, and then the resulting hash representation is compared against a list of known content. In the CSAM detection space, these hash lists of verified CSAM are stewarded by institutions like NCMEC and made available to technology platforms and Electronic Service Providers (ESPs) via API. Depending on the solution and the system used by the content moderators, this hashing and matching technology could support an age determination task. For example, if an image in the review queue is hashed and matches the content on a hash list of known CSAM, this content could be entered into a reviewing queue, indicating that the content depicts a child.

3. Image/Video Manipulation Detection Tools

Images and videos may be altered such that the individuals depicted in the content are made to look older or younger. These types of manipulations can be conducted using a variety of technologies, ranging from Photoshop to generative AI. The field of photo forensics has produced a range of solutions to authenticate content and detect content manipulation. These solutions are generally computer vision-based, including but not exclusive to AI.

4. Object Detection

Images and videos may contain contextual cues for the age of people depicted in the content (e.g., items that are typically purchased by a particular age demographic). Object detection and recognition consists of computer vision-based solutions that allow the location and classification of particular objects in an image or scene. For example, if an image contains a child-sized bicycle, this may be used as supporting evidence that the individual depicted in the image is a child.

5. Audio Detection

Videos may contain instances where people in the content utilize terminology or slang that is typically associated with a particular age demographic. Characteristics of the voices (e.g., vocal range) in the video may also be correlated with age. AI solutions (both those trained directly on audio and text-based natural language processing solutions) allow for the automatic detection of these types of keywords/phrases and vocal signals, which are then used as supporting evidence regarding the age of the person depicted in the video.

While separate from age determination tools typically applied to pieces of content, identity verification (generally used for age verification upon user sign-up for a product/service) may also benefit companies to make age determinations at a later stage in the content generation process.

6. Identity Verification Through Biometric Authentication and Identification

Biometric identification is a form of identity verification that relies on the use of distinguishing physiological characteristics to verify a person's identity. Biometric identification solutions can include the use of fingerprints, face embeddings, iris patterns, and other such markers. Biometric identification systems typically have three stages:

1. A user's biometric features will be extracted by the system
2. The system will compare these features against a pre-existing database of biometric identifiers and corresponding users
3. The system will return the user that corresponds with the set of provided biometric features (assuming that the user's biometric features are, in fact, already present in the database)

In this form of authentication, the user is making a claim that they are a particular person. As a result, the biometric system only needs to verify whether that is or is not the case by comparing the biometric features extracted against the biometric features of the particular person the user is claiming to be, as stored in the system's database.

Companies like Aylo require identity verification checks (via providers such as Yoti) for all uploaders and performers.

It is important to note that the use of technology is not foolproof, given its current accuracy levels. For example, Yoti's technology has greater accuracy for younger children than older children. For 6 to 12-year-olds, the technology has a mean absolute error of 1.3 years, and it is 1.4 years for 13 to 17-year-olds. Even with technology, therefore, it is difficult to distinguish between a 12-year-old and a 13-year-old or between a 17-year-old and an 18-year-old. This is also the case for human content moderation, as discussed below, and therefore, an approach that involves both technology and human review is a best practice, depending on the case and context at hand.

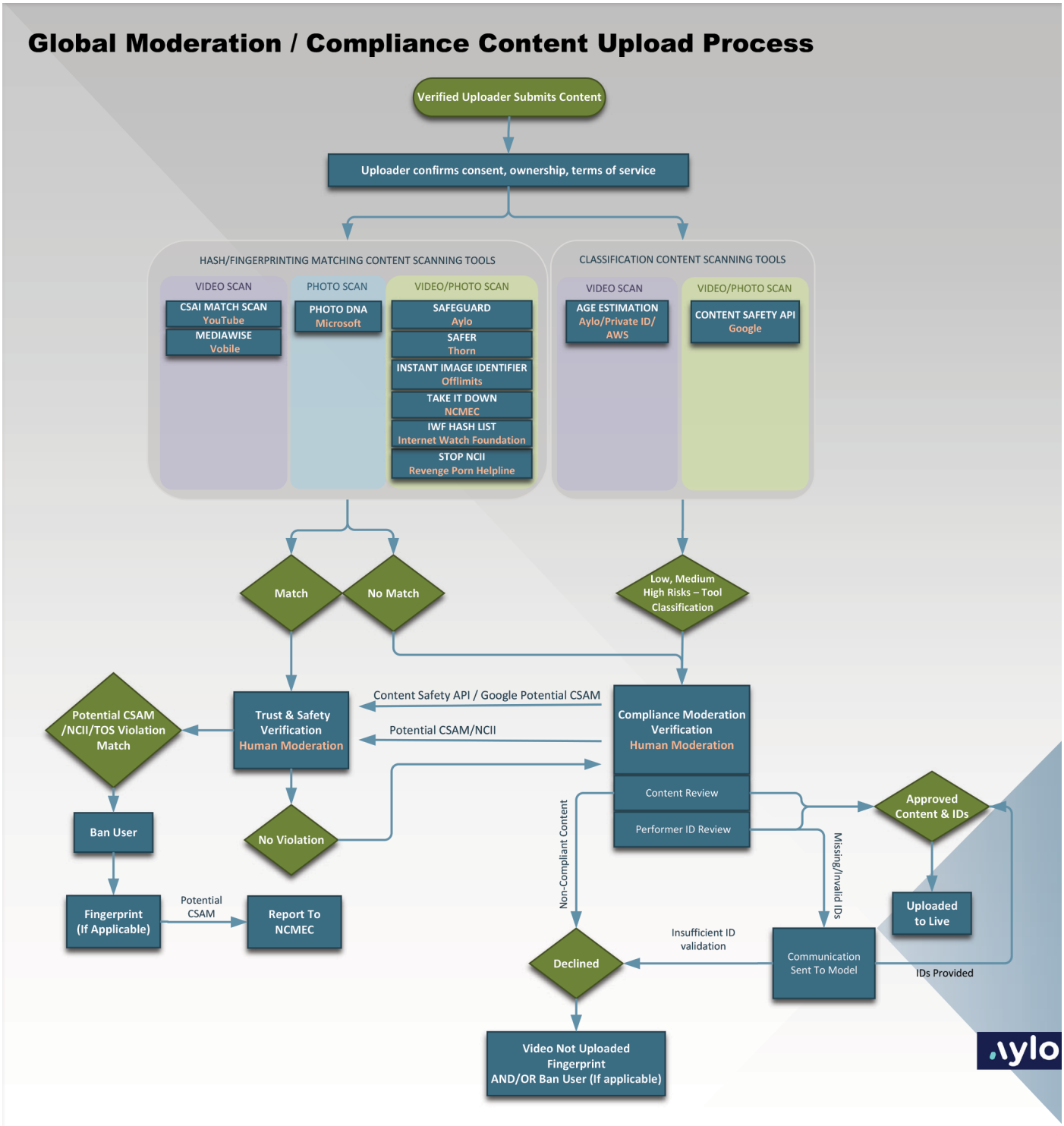
Section 3: Human Review to Support Age Determination

While technology can play a significant role in supporting age determination tasks, human review is currently crucial to confirm predictions made by AI or to review additional information that cannot easily be processed through AI at this stage. Triangulating information between AI predictions and human assessment can help make a more holistic and context-driven decision. The following areas are important to consider as they relate to human review for age determination by Trust & Safety teams within platforms or by content moderators within or outside platforms.

While consideration of the areas below may help improve the consistency and accuracy of age determinations, there will always be cases in which it may still be unclear as to the age of a person depicted or discussed in the content. In such instances, platforms should have clear guidance as to how this should be handled, erring on the side of caution.

Processes and Workflow

As discussed earlier, most companies use a combination of both technology and human review for age determination. Focusing on CSAM, since this is the most consequential area for accurate age determination, many companies build workflows/processes to optimize for detecting CSAM and then conduct a more nuanced age determination review if there is doubt as to whether a person in apparent CSAM is actually a child. Below is a sample workflow/moderation process for Aylo:



Like Aylo, companies can use a combination of technologies to cover the range of tech solutions previously described and to cover both new and existing content, as well as layer multiple tiers of human review (where necessary). Much of how this is structured depends on the size, resources, risk profile, and other such considerations of the platform.

Some AI classifiers (e.g., Google Content Safety API) give a confidence score when assessing age, which can help moderators in their determination or can be used to prioritize cases for further review. However, some companies state that confidence scores are never solely relied upon for a final age determination. While age determinations are required outside of the context of CSAM, there is not as clear and unique of a workflow for non-CSAM age determinations based on current knowledge and research.

Review Process

Technology helps in the moderator process flow because it can prioritize content for review. In some cases, it may be possible to use the confidence score produced by the automated AI solution to set expectations of content complexity. For example, if an age determination model has high confidence in its prediction for a particular piece of content, that may indicate that the content will require less time for a human to make a determination than manually reviewing the prediction and verifying or negating the automated recommendation. It is worth noting that with AI solutions, the "explainability" of a model's output can be difficult to ensure. While a high-confidence prediction could assist a moderator in making a faster determination, it may not be clear from the content why the model provided a high-confidence prediction, ultimately resulting in no change in the time a moderator takes to make an accurate age determination.

Using these confidence scores as a proxy for content complexity could allow platforms to set reasonable time expectations for moderators reviewing content. On this point, a best practice for moderation shared by Aylo could be that platforms have a cap —not a quota— for moderators being sent content on a per-unit basis to ensure a high level of accuracy.

Cultural Nuances and Trends

While more research in the field of digital anthropology is needed, there is research showing that certain races, ethnicities, or cultures may appear younger or older in online content to content moderators who are not members of that race, ethnicity, or culture. In addition to cultural norms, this could be due to physical differences in how people of different ethnicities age, as ethnicity and environment influence on tooth and bone development. A best practice to account for differences in apparent age due to race, ethnicity, or cultureand to reduce bias could be for platforms to apply existing scientific research to the development of age determination policies. Platforms can offer increased training and guidance to moderators based on digital anthropology trends such as:

- In markets/regions where there is a higher risk of content depicting children being uploaded to platforms
- For content determinations involving people whose races, ethnicities, or cultures have typically experienced over- or under-enforcement based on age-related policies with a goal of reducing moderator bias

- Where cultural nuances based on age range (i.e., typical slang by a certain age group) could be shared that may apply to text, image, or video content

Going back to the discussion from a technology perspective, a goal of moderating technology should be the reduction and elimination of bias on the basis of race, ethnicity, and culture. To build performant machine learning, the distribution of training data used in platforms' AI should reflect the end population of people depicted in the content, whether that means including full demographic distribution in the data set applied to a country or region or ensuring under-represented demographics are included. Training data sets should be representative of all population demographics in order to ensure accuracy and consistency. Testing of AI technology to ensure accuracy and fairness —focused on parity— is key to leveraging this technology effectively.

Policies

Platforms have a significant role to play in the effectiveness of age determination based on the initial policies that set the basis for moderation using relevant technology and/or people. Some of the key areas that need to be covered to establish well-thought-out policies on enforcing their community guidelines related to age determination include, beyond the minimum age of use for the platform:

- Minimum age of users depicted in content across all the various content types (e.g., dangerous acts, mature themes, etc.)
- Any content restrictions/limitations based on age (e.g., ad targeting); it may also be that the content is allowed but has boundaries on the monetization of the content or the algorithmic recommendation of such content for certain age groups
- Default product features (e.g., more restrictive privacy settings) for children (or younger age groups) versus adults

In addition to minimum age policies for the use of platforms, platforms may take it upon themselves to apply separate and additional age determination policies based on behavioral signals or other content (e.g., selfies posted by users) as users begin their activity on the platform, given that users may initially find a way to circumvent age gates.

Through discussion with child safety policy experts, moderators, and other subject matter experts in the field of age determination and digital anthropology, the following areas are highlighted as crucial areas for policy enforcement guidance to stakeholders involved in content moderation:

1. **Physical development guidelines and visual cues by age bracket:** This is used by moderators to make visual assessments (assisted or unassisted by technology) based on typical physical features by age. The Tanner scale is often used as a reference guide for moderators supporting platform moderation.
2. **Account-level information:** It can be extremely helpful for moderators to have the ability to look at additional account information that may provide relevant insights to make a more accurate age determination. It is important to note that this is not typically occurring with moderators who see content in a queue, outside of the platform's user-facing environment.

3. **Contextual cues:** There may be cues in images or videos that are typical to a specific age group. It is important that these types of contextual cues be documented by region, given cultural differences by market, so that trends can be gleaned and consolidated for future use.
4. **Signs of AI manipulation:** In addition to using relevant technology for determining AI manipulation, it is important that moderators also be trained and upskilled to spot signs of AI generation or manipulation. Although this is getting more difficult for people to do accurately, having an awareness of the trends in AI manipulation of content will be critical for catching content that may not have violated policy in its original form, but that does violate it based on its manipulation through the use of AI.
5. **Racial and ethnic differences:** As highlighted earlier in this paper, it is important to document and raise awareness of differences in noticeable physical development by age range, given that the perception of how old someone looks (and actual physical signs of aging) can vary according to race, ethnicity, or culture. Trends of under- or over-enforcement in markets should be well documented to minimize biases where possible.

Policy enforcement guidelines should also be clear as to how moderators should proceed if they are still unsure of an individual's age after following the relevant process/assessment guidance. Erring on the side of caution is often the norm, though platforms are continuously trying to reduce false positives, which are not only costly but could also result in real harm to users.

Training and Quality

Training is a key component to accurate age determination, and this training is critical to understanding and adhering to platforms' policies and guidelines. It is clear that well-trained moderators typically perform better on policy enforcement. Some training considerations include:

Period of Training:

While some platforms have a minimum training period (e.g., a certain number of weeks or months of product training, nesting, etc.), there is no standard amount of training for moderators across the Trust & Safety industry. Training will depend largely on the scope of the work, the number of systems/tools to be trained on, the size of the platform, the risk level inherent in the platform, the complexity of the policies involved, and other such factors. In general, moderators should be extensively trained until they can comfortably and objectively reach the accuracy that is required for the platform's needs. In general, there should be knowledge checks at different parts of the training period to assess the moderator's learning path and ensure that it is progressing according to the platform's needs.

Process and Content for Training:

Training moderators generally consists of an academic portion (e.g., understanding the type of content on the platform and learning the platform's policies) and an applied portion (e.g., practicing the application of policies in a sandbox environment) followed by nesting (where the moderator is operating in a live or production environment, but with additional help (e.g., more time to make a decision or greater assistance from a team lead and mentorship).

Given that the nature of the content requiring age determination can be sensitive, the focus of the training typically starts with an understanding of the policy, methodology, and systems for doing age determination, as well as the overall thought process/analytic framework that should be leveraged. Effective training typically gives moderators the ability to practically test and apply their knowledge during the training, as referenced briefly earlier in this section. It is important to emphasize that the content used in training should be vetted by platforms and should consider the role the moderator will have. At no time should CSAM be used for training purposes.

While there are a lot of factors impacting work-from-home vs. in-office requirements, for sensitive content review involving age determination, it may be beneficial to be able to have moderators work in the office, given that this can help provide a more controlled environment with a greater ability to receive in-person help from supervisors and other senior staff, peers, and in-person wellness support services.

Post-training Calibration and Review

After the initial training is completed, platforms should encourage ongoing process improvements to ensure moderators are consistently applying the policies, highlighting areas of lack of knowledge or policy ambiguities and staying atop the latest trends to ensure effective enforcement. A calibration process and a closed-loop feedback process help ensure guidelines are being met. Feedback on policies is reviewed, and moderators can be provided with specific coaching opportunities. Automated and/or human quality reviews can be leveraged for this purpose. In general, the following steps should be considered:

- A sample of cases should regularly go through a quality assurance process to ensure accuracy and other standards are being met across moderators.
- There should be a policy lead/QA manager who has created an age verification "golden set" (i.e., what is deemed the correct age determination on pieces of content) that is then reviewed against moderator determinations to capture and analyze any discrepancies.
- The quality assurance process should be used to help retrain and coach moderators and potentially provide feedback to inform platform policy updates or enforcement guidelines.

Section 4: Ecosystem Collaboration

Collaboration across the platform and Trust & Safety ecosystem is vital to addressing the challenges of ensuring children are not improperly depicted or monetized in content.

Collaboration between Industry, Government, and Civil Society

Recent laws and regulations worldwide now require increased child protection on platforms and increased moderating of content involving children. The protections required by these regulations for people under the age of 18 are not altogether consistent, however, and the practical application of countries' regulations of platforms is still developing.

In the absence of an overarching regulatory scheme governing child content on platforms, there have been productive recent efforts to collaborate on issues of child safety, from Project Lantern under the Tech Coalition to INHOPE's Universal Classification Schema to the work being done by the Digital Trust and Safety Partnership

Thorn, the Family Online Safety Institute, Internet Watch Foundation, International Centre for Missing and Exploited Children, and the Trust and Safety Professionals Association, among many other organizations. Despite these connections, there is not yet formal collaboration on the specific topic of age determination. This was highlighted by the lack of research, best practices, or content that was available to reference on age determination in the development of this report by the working group.

Now, with existing initiatives and organizations in place, it is the opportune time for partnerships to develop age determination best practices.

To advance collaboration on the topic of age determination, we recommend that there be a place to develop and share the latest trends, tools, and research on age determination, similar to what has been developed in other areas that are key to Trust & Safety. For sustainability and efficiency purposes, this could become a dedicated workstream or topic area of focus within existing, relevant NGO or industry-driven efforts. Industry players should be encouraged to share best practices in this regard and have an easily accessible forum to do so.

As part of the endeavor for this report, we hope that companies and other key stakeholders will be encouraged to track progress in this area and use this as a foundation to continue the development and sharing of best practices on age determination. Robust age determination practices will allow platforms to ensure the safety of content on their platforms for the benefit of their users and children across the world.

Authors:

- Bevans, Simone - Director of The Koons Family Institute on International Law & Policy, International Centre for Missing & Exploited Children
- Cooke, David - Sr. Director, Trust & Safety Regulations & Partnerships, Aylo
- Cunningham, Bob - CEO, International Centre for Missing and Exploited Children
- Lalani, Farah - Global VP, Head of Gaming, Trust & Safety Policy, Teleperformance
- Moyer, Alison - Corporate Counsel, Teleperformance
- Mudzongo, Rumbidzai - Legal Research Intern, International Centre for Missing and Exploited Children
- Dr. Portnoff, Rebecca Sorla - Vice President, Data Science, Thorn
- Pugalia, Akash - Former Global President, Media, Entertainment, Gaming and Trust & Safety, Teleperformance
- Sharma, Bindu - Vice President Global Policy & Industry Alliances, Managing Director, Asia Pacific, International Centre for Missing & Exploited Children
- Simpson, Michael - Staff Data Scientist at Thorn

- We thank Yun Choi (Sr. Director of Global Program Management, Teleperformance) and Lindsey Olson (Sr. Director of Global Operation, Teleperformance) for their support of this work.
- We also thank a number of Subject Matter Experts from various tech platforms and members of the Teleperformance Trust and Safety Advisory Council who reviewed/supported this work.

Teleperformance

Age Determination: *A Best Practice Guide for Online Platforms that Feature User-Generated Content*

May 2024