# Harmful, Hateful, and Deceptive/Misleading Content Guide

Version 1.1 – March 9, 2022

## 1. Introduction

In many types of evaluation tasks, you will be asked to identify different categories of content. This guide provides an overview of **Harmful, Hateful** and **Deceptive/Misleading** content, which includes content that can cause damage to individuals, groups, or society in one or more of the following ways:

1. **Harmful and Potentially Harmful Content**
    a. **Harmful to Self or Other Individuals:** Encourages, depicts, incites, or directly causes physical, mental, emotional, or financial harm to self or others
    b. **Hateful towards Specified Groups:** Promotes, condones, or incites hatred against a **Specified Group** of people
2. **Deceptive/Misleading Content:** Demonstrably misinforms people on important/sensitive topics, whether with the intent to deceive or while sincerely believing that inaccurate information is true

Although defined separately, these categories can overlap (e.g., content that is **Harmful or Potentially Harmful** can be **Deceptive/Misleading** as well).

Throughout this guide, you will see that some everyday terms are capitalized and bolded to emphasize that they have specific meanings. Your evaluation of **Harmful**, **Hateful**, and **Deceptive/Misleading** content should be based on the definitions and examples within this guide, not on your personal opinions, preferences, religious beliefs, or political views. Always use your best judgment and represent the cultural standards of your evaluation locale.

# 2. Harmful and Potentially Harmful Content

There are two subcategories of **Harmful and Potentially Harmful** content.

## 2.1 Harmful to Self or Other Individuals

This category of content is **Harmful and Potentially Harmful** because it encourages, depicts, incites, or directly causes harm to self or other individuals.

**Harm** includes physical, mental, emotional, or financial harm to people. Content should be considered **Harmful** if it directly attempts to harm people; encourages behavior that may result in harm; depicts extremely violent or gory content without a beneficial/educational purpose; or otherwise is severely traumatic to people who view the content.

Content does not have to be harmful to *all* people to be considered **Harmful**. Different people have different levels of vulnerability to scams, awareness of potential dangers (e.g., dangerous feats depicted in stunt videos), and tolerance for viewing violent/disturbing content. If there is a reasonable possibility that viewing particular content would cause **Harm** to those who are most vulnerable, it should be considered **Harmful.**

Content created with a beneficial purpose that reports on, discusses, or informs about harmful actions or events (e.g., fictional entertainment, reputable news, education) should typically NOT be considered **Harmful**. For example, advocacy aimed at drawing attention to harmful, real-world actions or events (such as content describing a protest against domestic violence) would not be considered **Harmful**.

Examples of **Harmful** content include:

- Content containing serious death threats or other realistic-sounding threats
- Content that shares personal information belonging to others with malicious intent to target them or promote harassment towards them (i.e., “doxxing”)
- How-to or step-by-step information that describes how to commit violent acts in an easy-to-replicate way
- Content meant to advocate for, glorify, or trivialize violence and atrocities, or to disparage victim(s) of violence/atrocities
- Content depicting or promoting information that facilitates or leads to serious harm to people or animals, or discussions that attempt to justify abuse of people or animals
- Content that encourages unsafe behavior or substantially downplays the risks of dangerous activities (e.g., consuming household cleaning products)
- Suicide promotion or pro-anorexia content that encourages people to engage in behavior that can result in hospitalization or death

- Health-related advice that contradicts well-established expert consensus and could result in serious harm (e.g., statements that lemons cure cancer), or could prevent someone from undertaking a life-saving treatment (e.g., encouraging a home remedy as a replacement for standard medical treatment to cure a disease)
- Malicious pages that attempt to scam or hurt people, contain suspicious links (e.g., to download malware), request personal information without a legitimate reason, or “phish” for passwords

Examples of content that should NOT be considered **Harmful** include:

- Depictions of violence in an action movie
- A news story about violent events
- Educational content that may depict violence or gross imagery
- An explanation of scams meant to raise awareness about them
- Portrayals of dangerous activities in a manner that does not promote the same feat (such as by clearly explaining the risks involved, describing the expertise and equipment required, etc.)

## 2.2 Hateful towards Specified Groups

This category of content is **Harmful and Potentially Harmful** because it promotes, condones, or incites hatred against a **Specified Group** of people.

For the purpose of identifying **Hateful** content, a **Specified Group** is a group of people that can be defined on the basis of:

- Age (e.g., older adults)
- Caste (e.g., Dalits)
- Disability (e.g., people who are blind)
- Ethnicity (e.g., Roma)
- Gender Identity and Expression (e.g., transgender people)
- Immigration Status (e.g., student visa holders)
- Nationality (e.g., Italians)
- Race (e.g., Asians)
- Religion (e.g., Christians)
- Sex/Gender (e.g., men)
- Sexual Orientation (e.g., lesbians)
- Veteran Status (e.g., Marines)
- Victims of a major violent event and their kin (e.g., victims of the Holocaust)
- Any other characteristic that is associated with systemic discrimination or marginalization (e.g., refugees, people experiencing homelessness)

Individuals (e.g., an actor/actress, a politician), membership-based groups (e.g., clubs, teams), and organizations (e.g., businesses, political parties) should NOT be considered **Specified Groups** for the purpose of identifying **Hateful** content.

Examples of **Hateful** content includes any content that:

- Encourages violence or ill treatment towards a **Specified Group**
- Promotes intolerance by demonstrating a staunch unwillingness to allow for the views, beliefs, or behavior of a **Specified Group**
- Implies that one **Specified Group** is superior or inferior to another
- Contains extremely offensive/dehumanizing stereotypes of a **Specified Group.** Note that stereotypes can be negative or positive.

The tone of the **Hateful** content must be *either* serious (i.e., not joking or satirical) *or* mean-spirited (i.e., with an intent to demean or promote intolerance) to be considered **Hateful**. Satirical comedy or artistic expression related to a **Specified Group** should NOT be considered **Hateful** unless it is clearly mean-spirited.

Criticism of objects, philosophies, and ideas are generally NOT considered targets of **Hateful** content. For example, negative criticism of a religious doctrine should NOT be considered targeted at the **Specified Group** that follows that religion. Remember that the content must promote, condone, or incite hatred of *people* to be considered **Hateful**.

Educational content (e.g., definitions, research, academic papers), news stories, or other content that has a beneficial purpose of informing society about or exposing hate-related issues/topics/incidents should NOT be considered **Hateful**. Similarly, historical documents/videos that aim to capture the beliefs of different eras should NOT be considered **Hateful**.

Examples of content that should NOT be considered **Hateful** include:

- A historical documentary of WWII featuring speeches from Nazi leaders
- A stand-up comedy routine that plays off of stereotypes in a way that is not mean-spirited
- A newspaper article about a hate organization
- The dictionary definition of a slur
- A discussion about a particular religious text and its views on women

## 3. Deceptive/Misleading Content

This category of content misinforms people on important/sensitive topics in ways that can cause harm to people and society.

**Deceptive/Misleading** content may have been produced with the intent to misinform people, or when the content creator may believe that the inaccurate information they are sharing is true. There is an especially high standard for accuracy on important/sensitive topics that could potentially impact a person's life or affect society's ability to maintain an informed citizenry, including topics such as:

- News and current events (e.g., international events, politics, science, etc.)
- Civics, government, and law  (e.g., information about voting, government agencies, social services, etc.)
- Finance (e.g., information regarding taxes, insurance, banking, etc.)
- Health and safety (e.g., advice or information on medical issues, nutrition, emergency preparedness, etc.)

Please be more sensitive to **Deceptive/Misleading** content that involves important/sensitive topics, and be sure to research consequential facts or claims if you aren't sure.

Content should be considered **Deceptive/Misleading** when it contains *at least one of the following:*

- **Clearly inaccurate information** that can easily be refuted by simple straightforward facts (e.g., false claims that a world leader has died, misleading or false statistics on gun violence, etc.)
- **Claims that contradicts well-established expert consensus** (e.g., claims that lemons cure cancer, content denying that climate change is real), with expert consensus defined as a set of positions, facts, or findings that are widely agreed upon by authorities in the relevant field (e.g., widely-adopted medical guidelines, an investigative report put forth by a relevant watchdog group, etc.)
- **Unsubstantiated theories/claims** not grounded in any reasonable facts or evidence, especially those that could erode confidence in public institutions. This includes unsubstantiated theories that have either been thoroughly debunked (e.g., the 9/11 attacks were planned by the United States government) or are too outlandish to be given credence (e.g., several world leaders are lizard people).

However, note that some types of information are subjective, debatable, unverifiable, or inconsequential. For example, content should NOT be considered **Deceptive/Misleading** if it *exclusively* contains:

- Content created with a clear entertainment purpose, containing no hard claims of factual accuracy and with no damage to people or society. Examples include many types of fiction, satire or parody, astrology, folklore, myths, and urban legends.
- Reviews expressing personal preferences, opinions, or value-based judgments about a product, restaurant, book/movie/TV show, etc.
- Claims or statements that are reasonably debatable when there is not a single established correct answer or truth (e.g., discussions about the relative effectiveness of different healthcare systems)
- Insignificant errors or inaccurate information about a trivial topic (e.g., inaccuracies in the height of a celebrity)

Content that aims to persuade others that a certain position or perspective is correct is fairly common on the Internet. One-sided/opinionated/controversial/polarizing content should NOT be considered **Deceptive/Misleading** unless it is **Harmful or Potentially Harmful** (as described above) and contains clearly inaccurate information, contradicts well-established expert consensus, or is not grounded by reasonable facts/evidence.

In certain cases, the **Deceptive/Misleading** nature of content may not appear directly in the video itself. Instead it could involve the content creator (e.g., a creator blatantly misrepresenting their medical credentials for a video on medical topics), or the purpose of the content (e.g., a video title is deceptive/misleading, even if the video content on its own is not).

Finally, note that **Deceptive/Misleading** content can be especially hard to identify because it may require research from outside sources. Reputable fact-checking websites can't always keep up with the volume of unsubstantiated theories/claims produced by the Internet, and some theories may even claim that debunking information is inaccurate. You should attempt to find high-quality, trustworthy sources to check accuracy and seek out the consensus of experts if you are unsure, particularly for important/sensitive content. Please research theories and claims to the extent the task time allows. If a theory/claim seems wildly improbable and can't be verified by independent trustworthy sources, you should consider it unsubstantiated.

## 4. Harmful, Hateful, and Deceptive/Misleading FAQ

| **Question** | **Answer** |
|---|---|
| 1. Can content fall into more than one of these categories? | Although Harmful, Hateful, and Deceptive/Misleading categories have been defined separately above, content may fall into more than one of these categories. For example:<br>● Content that spreads hateful misinformation regarding a **Specified Group** is both **Hateful towards Specified Groups** and **Deceptive/Misleading**, such as claims that a **Specified Group** is solely responsible for a financial crisis<br>● Content that encourages violence towards a **Specified Group** is both **Hateful towards Specified Groups** and **Harmful to Self or Other Individuals**<br>● Medical information that contradicts well-established expert consensus and could result in serious harm is both **Deceptive/Misleading** and **Harmful to Self or Other Individuals** |
| 2. What is the threshold for content to be considered **Harmful**, **Hateful**, and/or **Deceptive/Misleading**? | It will sometimes be difficult to determine whether specific content meets the threshold to be considered **Harmful**, **Hateful**, and/or **Deceptive/Misleading**. There is lots of content on the Internet that some would find controversial, one-sided, off-putting, or distasteful, yet would not be considered **Harmful**, **Hateful**, and/or **Deceptive/Misleading** based on the definitions in this guide. The distinguishing feature of this content is that it is **Harmful**, **Hateful**, and/or **Deceptive/Misleading** to a degree that can be damaging to individuals or society as a whole. |

| Question | Answer |
|---|---|
| 3. What video page details should I take into account when evaluating **Harmful**, **Hateful**, and/or **Deceptive/Misleading** content? | Your evaluation should be based primarily on the content in the task (video content, thumbnail, title, description) and its purpose.<br>Your evaluation should not take into account the number of views, likes, or subscribers. Likewise, issues found in ads or supplementary content outside of the creator's control (e.g., platform-generated recommendations to similar videos, user-generated comments) can be excluded from your evaluation. |
| 4. How should I take the reputation of the creator/channel into account when evaluating **Harmful**, **Hateful**, and/or **Deceptive/Misleading** content? | A creator who has a controversial reputation in other areas, or even the topic at hand, can create content that is not necessarily **Harmful**, **Hateful**, and/or **Deceptive/Misleading** content. However, a creator/channel's overall reputation may provide useful context.<br>By reviewing other videos or information from the same creator/channel or information about them from the broader web, you may be able to determine their credibility or potential expertise on a topic discussed in a video. For instance, if a video seems to be about an unsubstantiated claim but by looking at the channel you find that this channel is a well known fact checker, that may help you assess that the video is likely debunking an unsubstantiated claim, not promoting it. |