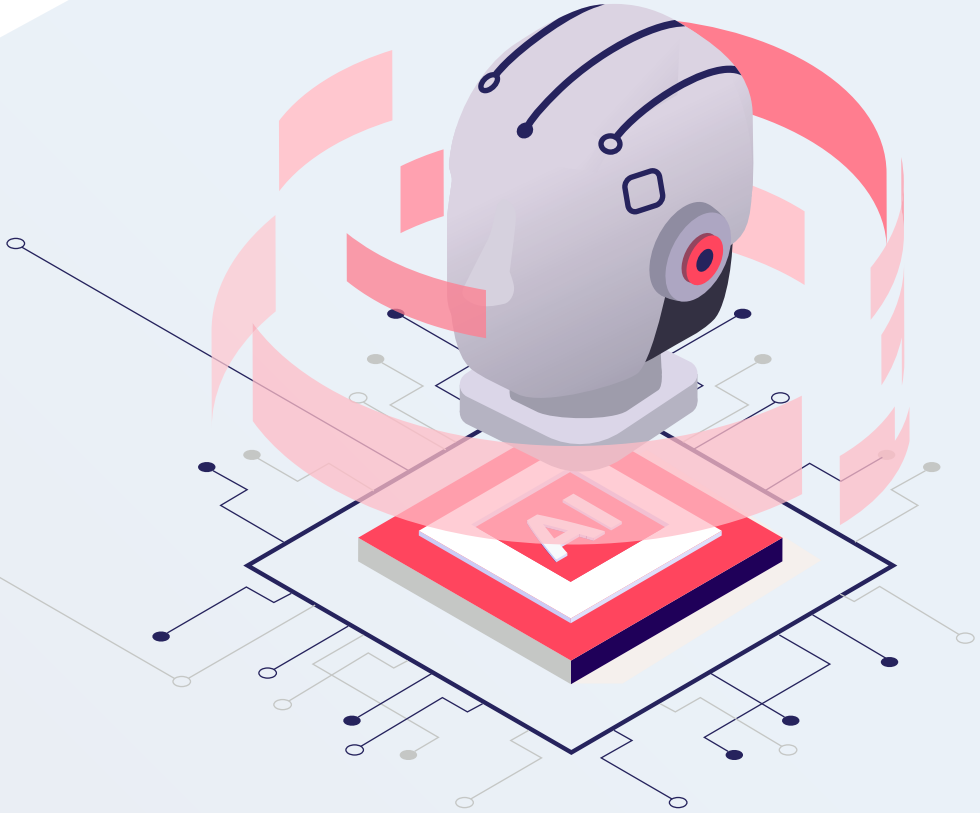
SUTHERLAND



White Paper

THE NEW CONTENT SAFETY PARADIGM: IMPEDIMENTS AND COMPETENCIES OF GEN AI IN CONTENT MODERATION

The growth of online platforms, content, and AI capabilities is creating both opportunities and challenges for content safety. Learn how your business can effectively moderate content in this rapidly evolving ecosystem by harnessing AI-human synergy.

TABLE OF CONTENTS

INTRODUCTION

The world continues to witness an **unprecedented rise in data generation** as internet connectivity, social media, smart devices, and digital platforms continue to proliferate and become an integral part of our daily lives. Activities such as online shopping, social media usage, media streaming, online education, online banking, and online conversations are now ubiquitous.

As a result, along with individual users, businesses and public authorities are also among the sources of ever-growing data. The internet **has democratized content creation**, allowing billions of users to share their thoughts, ideas, and creativity with a global audience.

*Online Data Created Has **Increased Exponentially***

*Approximately **328.77 million** terabytes of data are created each day*

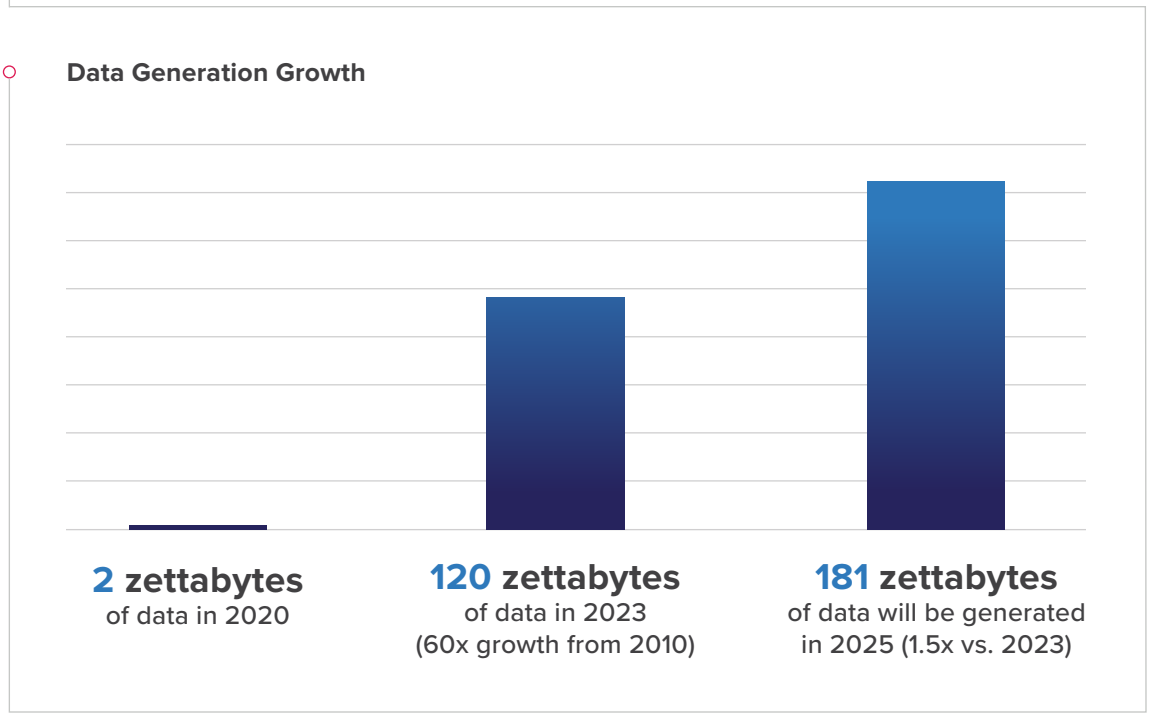


Illustration 1: Global Data Generation Growth[1]

1 Source: https://explodingtopics.com/blog/data-generated-per-day

SOCIAL MEDIA: THE PRIME PROPONENT OF USER-GENERATED CONTENT (UGC)

Social media forms a significant component of digital engagement in today's age, and has therefore seen rapid adoption globally as access to fast internet continues to become a part of daily lives everywhere.

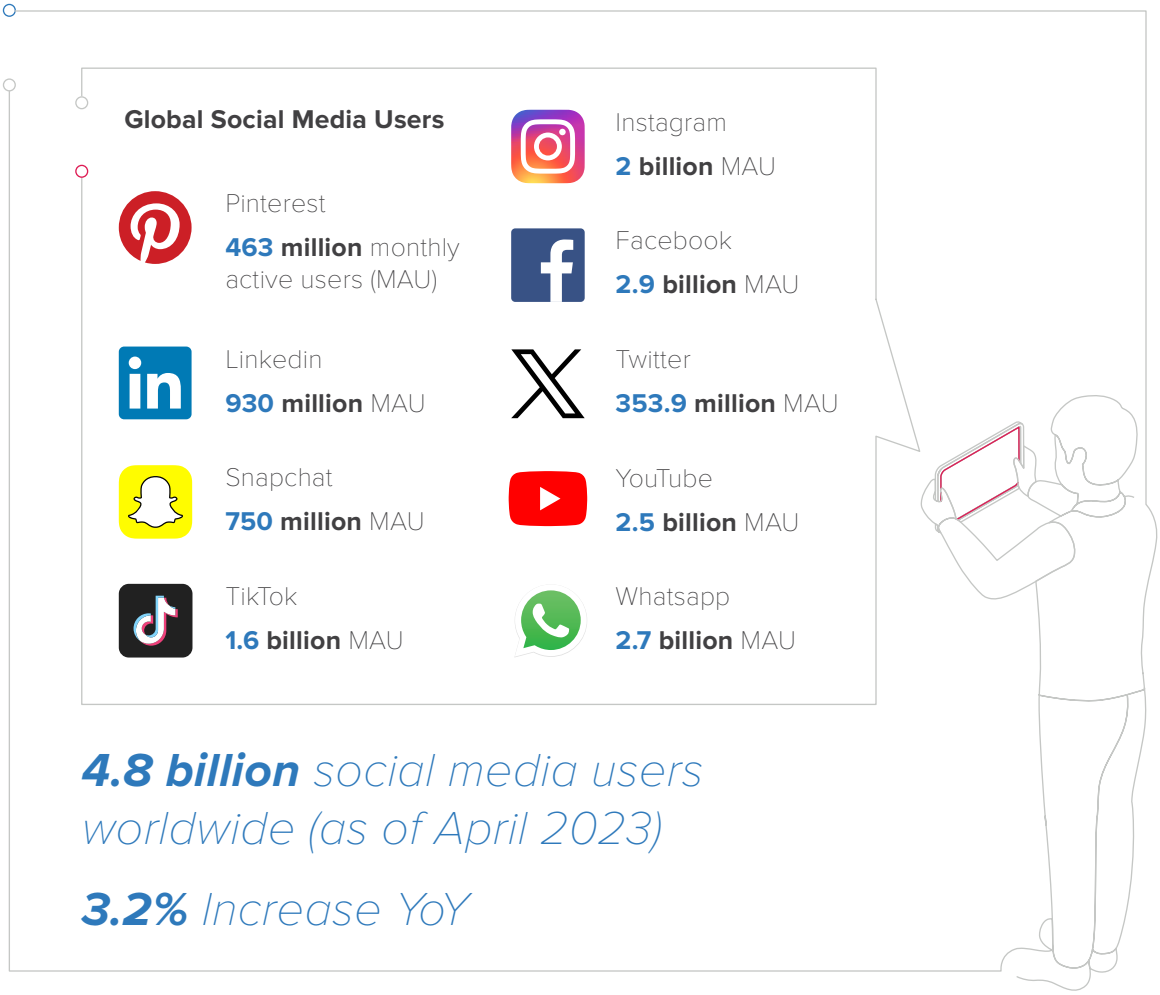


Illustration 2: Social Media Users - Global[2]

The growing penetration of social media and digital platforms has caused user engagement on online platforms to rise dramatically in the last decade. Online platforms have become **an integral part of global communications** – and now perhaps supersede other media as the consumer preference for personal communication.

2 Source: https://www.searchenginejournal.com/social-media-statistics/480507/#close

Growth of Video

Higher propensity and preference for online engagement has also meant video-based content such as live streams and ephemeral content – such as short-form videos and stories – have grown rapidly in popularity:

66% ***of online users***

in 2022 found short-form videos to be the most engaging type of content (vs. 50% in 2020).[3]

73% ***of consumers***

prefer short-form videos to search for products or services.

At the same time, streaming video has seen a rise globally:

23% ***of global viewing time***

is accounted for by live video streams

41% ***of internet users***

watch live streams.[4]

163.4M ***consumers***

of live-streamed content in the US alone in 2023.[5]

3 Source: https://www.linkedin.com/pulse/essential-social-media-video-metrics-marketers-should

4 Source: https://www.uscreen.tv/blog/live-streaming-statistics/#:~:text=Data%20shows%20that%20live%20streaming,have%20watched%20a%20live%20stream.

5 Source: https://www.demandsage.com/live-streaming-statistics/

GEN AI IS A **DOUBLE-EDGED SWORD**

While users and organizations ride on the digital engagement wave, the advent of **AI-enabled capabilities** have also had a significant impact on digital content.

The AI Paradox

AI is simultaneously proliferating digital content and helping the content moderation process

Bright side - AI can help with Content Moderation and Trust & Safety processes

- Combating time sensitive content (ephemeral and livestream content)
- Content Moderation at Scale
- Proactive Automated Content Moderation
- Detect and preempt child grooming at early stages
- Swift Identity verification leveraging multiple data points
- Protect Employee Wellbeing

Dark side - Generative AI can proliferate undesirable content

- Fake/ Scam Bots
- Disinformation
- Deep Fakes
- Hallucinations

Content is growing by leaps and bounds

Illustration 3: AI – The double-edged sword

Rapidly Evolving Challenges

AI has become widely embraced, bolstered content generation capabilities, and has flooded the internet with substantial volumes of content. Platforms such as OpenAI's ChatGPT and DALL-E enable users to **generate high-quality text and images quickly and easily.** While this freedom of the digital age has unlocked countless positive opportunities for expression and connection, it has also given rise to numerous challenges including AI hallucinations, deep fakes, and both misinformation and disinformation.

According to a report by the Financial Times, the number of **deepfakes used in scams** in just the first three months of 2023 were higher than the entirety of 2022.[6]

Among the strongest examples of deepfakes in the media are the **spread of fabricated videos** of Hillary Clinton and Joe Biden in the run up to the 2024 presidential race, which are among many such videos flooding social media – potentially having a deep impact across people navigating the polarized US political landscape.[7]

In April 2023, Newsguard – a company working on tracking and combating misinformation and disinformation online – identified **49 news and information sites almost entirely generated using AI.**[8]

6 Source: https://www.ft.com/content/167befa0-123f-4384-a37e-c8a5b78604b2

7 Source: https://www.reuters.com/world/us/deepfaking-it-americas-2024-election-collides-with-ai-boom-2023-05-30/

8 Source: https://www.newsguardtech.com/special-reports/newsbots-ai-generated-news-websites-proliferating/

A Rising Tide of Online Victimization

These AI-driven capabilities and growing volumes of data have also allowed bad actors to find more ways to engage in cyberbullying by uploading **objectionable, offensive, harmful, fraudulent, abusive, and misleading content online:**

- Close to half of teens online face some sort of cyberbullying[9]
- **38% of users** experience cyberbullying on social media platforms daily[10]
- Marginalized groups face increased harassment online:[11]
  - **76% of users** that identify as transgender report online harassment
  - **38% of respondents** that identify as Black and African American face some form of digital abuse regularly

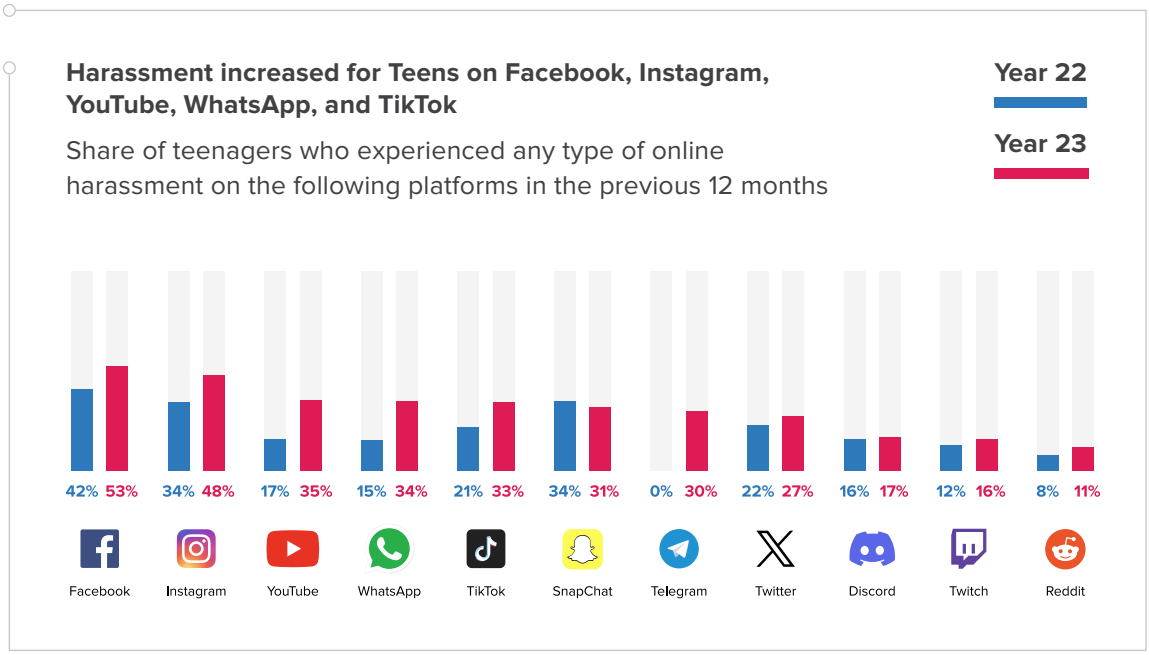


Illustration 4: Share of Teenagers (US) facing Online Harassment[12]

The burgeoning number of threats will require greater responsibility from **digital platforms and the Guardians of the Digital World** (the Trust and Safety or Content Moderation mechanisms). The need for effective content moderation and content integrity has never been more crucial in today's digital age.

9 Source: https://www.comparitech.com/internet-providers/cyberbullying-statistics/

10 Source: https://www.pandasecurity.com/en/mediacenter/family-safety/cyberbullying-statistics/

11 Source: https://in.mashable.com/digital-culture/55526/more-than-half-of-americans-have-experienced-online-harassment-says-adl-report

12 Source: https://in.mashable.com/digital-culture/55526/more-than-half-of-americans-have-experienced-online-harassment-says-adl-report
Source: YouGov Survey on behalf of ADL. *2023 was the first year the survey asked teens about Telegram.

RISING TO THE CHALLENGE: COMBATING CONTENT AT SCALE

As online content grows in volume, it becomes difficult for platforms to combat the challenges of harmful and inappropriate content with traditional, manual moderation methods which rely on humans to screen and filter out content that is deemed unfit.

Humans simply aren't equipped to consume content at scale and make consistent decisions to filter it out – all while mitigating the psychological impact of viewing and reviewing monumental levels of such content.

Platforms therefore increasingly rely on the capabilities of Artificial Intelligence (AI) for content moderation. AI-based moderation provides the following crucial capabilities to counter the rising tide of UGC.

Scale and Efficiency

AI content moderation solutions can process vast quantities of text, images, and videos in real-time, and at unprecedented speed – allowing platforms to review and moderate (or flag content for human moderation) quickly and accurately.

Consistency

AI can make decisions independent of any biases that human moderators may have, and is also immune to other factors such as fatigue, tiredness, monotony, and emotional connection which lead to inconsistencies in content moderation decisions.

Round-the-Clock Availability

AI-powered content moderation operates around the clock, providing constant vigilance against inappropriate content – a necessity in a globally operative digital ecosystem.

DECIPHERING GEN AI'S PROWESS IN TACKLING CONTENT NUANCES

Advancements in technology and AI training has meant that it is keeping up, to a certain extent, with the ever-changing nature of human conversations in the digital space. With bad actors continuously changing tactics and devising new ways to evade traditional content filters, AI systems are adapting rapidly to confront these evolving challenges.

Machine learning algorithms are being trained to recognize emerging trends and adapt moderation techniques accordingly – a crucial capability in addressing novel threats such as deepfakes, hate speech, and misinformation.

Language Understanding

AI models can comprehend and interpret text in multiple languages, identifying hate speech, abusive language, and other harmful content across linguistic barriers. Training on trends (with closed loop feedback from frontline moderators) can additionally enable AI models to identify trending slang and slurs. This makes AI content moderation a valuable tool for global platforms that span diverse geographies and cultures.

Image and Video Recognition

As one of the most sought-after capabilities of the technology, AI-driven image and video recognition is utilized heavily to screen explicit or graphic material to reduce its distribution on digital platforms. It also reduces the exposure of human moderators to highly disturbing content by blurring and other masking techniques.

Real-Time Moderation

AI can significantly bolster capabilities for analyzing live content. Growth in live and streaming media has meant that moderating real-time data is crucial in ensuring user safety. AI can instantly and automatically detect harmful cases before live streaming.

Contextual Understanding

AI content moderation models are becoming ever more adept at understanding context, differentiating between satire, humor, and genuinely harmful content. However, AI still requires a human-in-the-loop to ensure legitimate content remains accessible.

SYMBIOSIS IN CONTENT MODERATION: GEN AI AND THE HUMAN-IN-THE-LOOP

While AI has rapidly evolved into a powerful tool for content moderation, it cannot replace the value and holistic impact of human oversight. Human moderation is still essential, and is necessary in parallel with automation to address complex moderation cases where finer nuances of human nature and intent are at play, as well as in refining AI algorithms.

Collaboration between AI and human moderators is an effective way to address the challenges of content moderation in online platforms. AI and humans can work in unison and build a synergistic mechanism that combines the efficiency of AI with the nuanced judgment of human moderators.

Embracing AI-Human Synergy

Artificial Intelligence (AI)

Scalability: AI enables handling of high volumes of content, allowing human moderators to focus on the most critcal and nuanced cases.

Initial Screening: AI algorithms can be used for initial content screening - involves the automatic detection of potentially harmful, disturbing or inappropriate content, avoiding exposure to human moderators.

Pattern Recognition: Scan large volumes quickly to recognize patterns and trends in UGC to apprise moderation policy and guideline refining.

Reducing Psychological Impact on Moderators: AI can filter out the most explicit and harmful content, reducing the exposure of disturbing content to human moderators.

Human Moderator

Handling Complex Cases: Evaluating content that appears harmful on the surface but is actually not when finer nuances of the context are judged (and vice versa).

Cultural and Context: Also linked with complex case, humans are better at understanding cultural nuances, slang, which may be challenging for AI. This also pertain to training algorithms to interpret accurately.

Evolving Guidelines: Setting and adapting moderation policies and guidelines in accordance with patterns and trends.

Training of AI Algorithms: Human moderators can help train AI systems by providing feedback on false positives and false negatives.

24/7 Coverage: Combined together, AI and human moderators can provide continuous content moderation, including during off-hours for human moderators.

Illustration 5: AI-Human Moderation Synergy

Safeguarding the Guardians of the Internet

Apart from augmenting moderation capabilities, AI can also play a crucial role in mitigating the psychological effects harmful content can have on human moderators. This is possible by leveraging AI across operational and wellness perspectives within content moderation practices.

Screening Out Harmful/Egregious Content

AI algorithms automatically filter out harmful or egregious content.

Reducing Burden of Scale and Complexity

AI can help in pre-moderation – such as conducting first-level filtering, sentiment analysis, and categorization of content – to reduce the burden of multiple decisions that human moderators may need to make for several thousand pieces of content daily. Here, AI can help with pattern recognition, historical analysis, and contextual analysis (to a limited scale) to enable faster decision making for humans.

Supporting Mental Health

Proactive wellness is more of a necessity than a perk in today's content moderation scenario, and AI can go a long way in monitoring moderators' interactions and behavior to preemptively flag instances of possible distress. This also enables teams to put early warning systems in place, and install pre-emptive wellness practices.

SUTHERLAND'S GEN AI PEDIGREE AND CONTENT MODERATION PRACTICE

While AI has rapidly evolved into a powerful tool for content moderation, it cannot replace the value and holistic impact of human oversight. Human moderation is still essential, and is necessary in parallel with automation to address complex moderation cases where finer nuances of human nature and intent are at play, as well as in refining AI algorithms.

Collaboration between AI and human moderators is an effective way to address the challenges of content moderation in online platforms. AI and humans can work in unison and build a synergistic mechanism that combines the efficiency of AI with the nuanced judgment of human moderators.

Contact Us to learn more about how Sutherland can add value to your content safety and customer experience strategy.

We make digital **human**™

sutherlandglobal.com
sales@sutherlandglobal.com
1.585.498.2042

Sutherland is an experience-led digital transformation company. Our mission is to deliver exceptionally designed and engineered experiences for customers and employees. For over 35 years, we have cared for our client's customers, delivering measurable results and accelerating growth. Our proprietary, AI-based products and platforms are built using robust IP and automation. We are a team of global professionals, operationally effective, culturally meshed, and committed to our clients and to one another. We call it One Sutherland.

SUTHERLAND®