R‘AI’sing the Bar: The Enhanced Role of AI in Ensuring Safe Digital Spaces in the Future

Enhance your content moderation services with 5 practical AI-based insights covering all the bases from initiation to augmentation

Wipro advocates the findings in this report by Everest Group

Everest Group®

R‘AI’sing the Bar: The Enhanced Role of AI in Ensuring Safe Digital Spaces in the Future

Manu Aggarwal, Partner
Abhijnan Dasgupta, Practice Director
Darshita Lohiya, Senior Analyst
Aseem Nousher, Senior Analyst

Contents

Introduction

From automation to process augmentation, from evolving search capabilities to enhanced user experience, Artificial Intelligence (AI) has taken the world by storm. It has hyper-scaled digital content, made easier by increasing internet penetration, and created new possibilities in content generation, ad creation, digital marketing, and personalization. According to experts, by 2025, over 90% of all digital content could be generated by AI.[1] Generative AI (GAI) platform ChatGPT has already made history by reaching 100 million monthly active users just two months after launch.[2]

However, as with all new technologies, there are pros and cons – AI has created new challenges and exacerbated existing ones.

Technologies such as GAI pose numerous threats to the digital space including misinformation, plagiarism, lack of governing regulations, proliferation of unacceptable/dangerous content, lack of context, biases, and lack of transparency in AI algorithms, among others.

Safety has always been a priority among all stakeholders including platforms, providers, regulators, and the users themselves. To ensure safe spaces, platforms currently use a combination of humans and technologies to moderate digital activity. However, vast volume and content complexity, rapid digital evolution – including the creation of Web3-enabled virtual worlds and technologies like GAI – have challenged existing moderation methodologies.

Platforms and providers use various AI-enabled technologies such as Machine Learning (ML) and Natural Language Processing (NLP) for advanced analytics, content moderation, user authentication, and fraud prevention. However, AI’s potential to handle the entire universe of safety concerns is yet to be explored. The advent of GAI calls for a shift in the way we approach content moderation today, compelling increased investment to develop a robust AI infrastructure.

In this report, we explore the current state of technologies in the Trust and Safety (T&S) landscape and the vast potential of AI in augmenting user safety. We propose an AI investment strategy for enterprises based on a holistic platform-based approach that encompasses both the right tools and solutions and responsible AI principles to ensure ethical and sustainable moderation practices. Finally, we make an argument for stakeholders to adopt a collaborative mindset to reach full scale potential in creating a robust AI infrastructure for a safe digital space.

1 Yahoo Finance
2 Similarweb

A safe digital space is fundamental to a holistic user experience

Everest Group take

A safe digital space is characterized by personalization, transparency in policies, protection of user privacy, and proactive content moderation. Lack of adequate content moderation facilities will result in poor customer experience, damaged brand reputation, and customers and advertiser churn.

The digital space has evolved from simple text-exchange forums to multi-media platforms that host a variety of content formats ranging from text and videos, through live streams, to 3D digital worlds. These platforms are the cornerstone of communication, entertainment, and financial transactions, among other things. They have become integral to how people interact, function, learn, think, and shop.

With increasing penetration of these platforms in day-to-day lives, digital safety has shifted from nice to have to fundamental to the holistic user experience.

A **safe digital space** is a protected ecosystem that ensures the safety and interests of all stakeholders including users, platforms, regulators, and society in general. Some characteristics of a safe digital space include:

- Collaborative and healthy exchange of ideas, information, and opinions
- Protected user privacy, identity, information, and financial transactions
- Personalized user experience through localization and language and cultural inclusivity
- Transparent policies and procedures, and clarity around potential risks to user safety
- Proactive moderation of all forms of content and removal of misinformation and harmful content
- Adherence to regulations on digital safety

The absence of these core elements in a digital ecosystem reflects negatively on platforms operating in it and erodes trust among users, regulators, and other stakeholders. As platforms strive to enhance user experience to differentiate themselves in the competitive digital landscape, they need to prioritize holistic user safety to build trust and retain users.

However, the expansion of the digital landscape and the content within it has created specific impediments to the development and maintenance of safe digital spaces:

- Increasingly non-distinguishable AI and other auto-generated data
- Proliferation of toxic content such as hate speech, violence, pornography, and Child Sexual Abuse Material (CSAM)
- Data and information breaches, which compromise user identities
- Identity theft and financial fraud
- Dynamic regulatory landscape creating non-uniform platform policies around content moderation, information collection, and user safety

To bridge the gap between the current and ideal/safe digital space, a scalable, agile, and industry- and use case- agnostic solution that can facilitate proactive moderation of diverse content is needed. Advanced AI technologies can help detect and prevent threats and vulnerabilities that compromise the safety of digital platforms. Additionally, AI can stimulate transparency, privacy, and sustainability in enterprise T&S policies.

Invest in intelligent solutions to ensure a safe digital ecosystem

Everest Group take

There are multiple AI technologies in use today that help in increasing the efficiency of T&S operations; emerging technologies such as GAI have the potential to disrupt the T&S market. While GAI could cause a surge in toxic content on the internet, increasing the need for moderation, it can also be used to create agile and scalable T&S solutions. AI solutions hold tremendous potential to enhance T&S operations by overcoming their inherent challenges and augmenting human intervention.

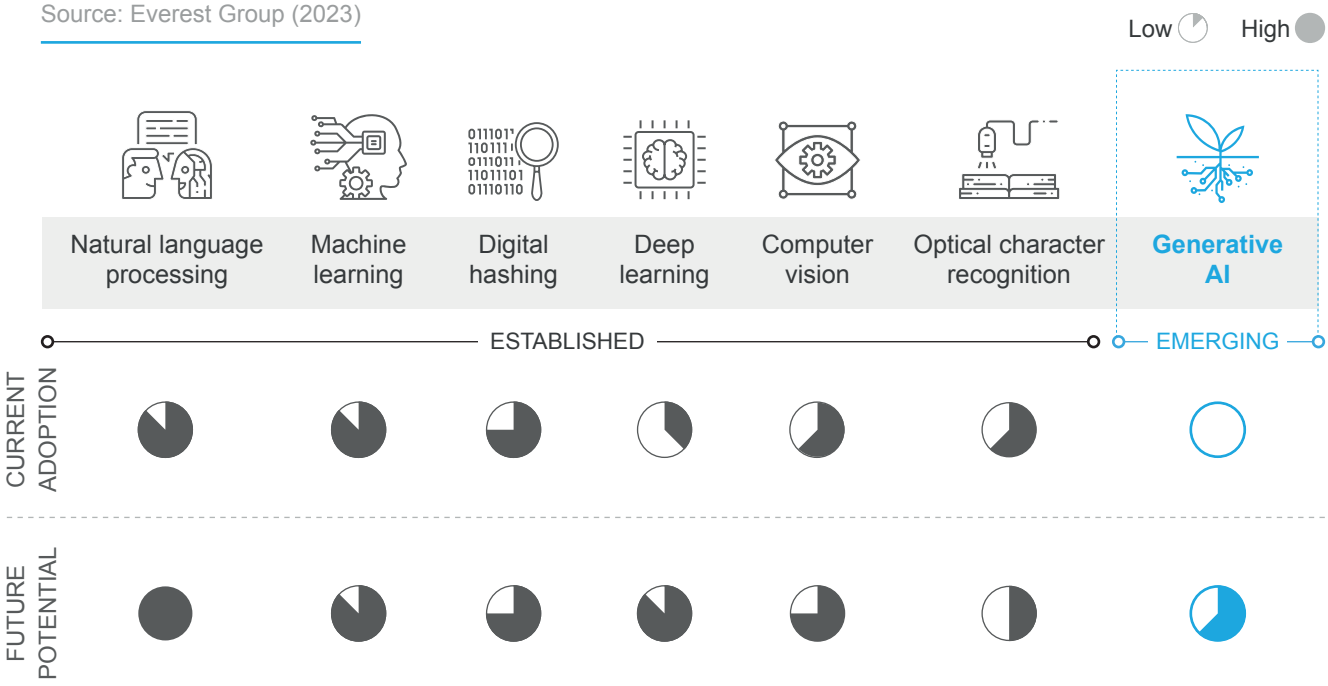
Current AI technologies address multiple T&S use cases and improve operational efficiency

In addition to boosting content creation, AI technologies play a vital role in T&S by providing operational scalability and agility. Intelligent moderation tools can review large and diverse content sets in real time, proactively remove harmful content, and analyze behavioral patterns to detect anomalies. They can further prevent cyber-attacks by monitoring security breaches and enable risk and demand forecasting based on historical data analysis and projections. Exhibit 1 shows key intelligent technologies in T&S and measures their current adoption and future potential.

EXHIBIT 1

Intelligent technologies in T&S

Source: Everest Group (2023)



NLP and ML are the most widely used technologies that facilitate proactive content moderation through text and sentiment analysis and pattern recognition. Deep learning can augment ML to increase the accuracy of content moderation, risk assessment, and fraud detection. Digital hashing is used to identify and remove explicit content such as CSAM and its digital copies. Enterprises, especially banks and other financial institutions, use Optical Character Recognition (OCR) to scan text embedded in images to detect fake user identity proofs and financial documents. Computer vision enables face recognition to moderate live streams and examine images and videos for harmful content.

With increasing awareness around mental health, platforms are deploying various intelligent tools and solutions to reduce moderator exposure to harmful content and preserve their overall wellbeing. Organizations are using automated chatbots powered by conversational AI to gauge moderator moods, and wearable devices track fluctuations in pulse, brain waves, and body heat in response to violent and graphic content. Computer vision can perform the first level of content screening to detect and remove egregious content before it reaches human moderators thus limiting their exposure and reducing overall workload, especially during live moderation. Additionally, technologies such as ML and deep learning can be used to triage and queue content based on different levels of toxicity and urgency and ensure that the tasks are assigned to the moderators with the right skill sets.

The scope of AI technologies in T&S extends beyond social media to various industries and work types including user profiling in gaming and dating, increasing search relevance in e-commerce, and real-time moderation for online streaming. Exhibit 2 offers several industry use cases for AI in T&S.

EXHIBIT 2

Industry use cases for AI in T&S

Source: Everest Group (2023)

ONLINE DATING

Online dating apps use NLP to combat hate speech and offensive messages, computer vision to identify and prevent fake profiles and catfishing, and deep learning algorithms for nuanced user profiling and better matchmaking.

CASE-IN-POINT

Tinder users get an "Are You Sure" prompt before sending messages that its AI algorithms detect as harmful or offensive.

ONLINE SHOPPING

AI models can identify fake user profiles and proactively remove fake reviews to prevent online fraud and establish product and seller credibility.

CASE-IN-POINT

Mozilla acquired Fakespot, an AI-based startup that detects fake reviews and scams on online shopping platforms. The service will be provided as a browser extension of Firefox.

LIVESTREAMING

AI technologies facilitate real-time moderation of text, audio, and visual content in multiple languages, which protects users against cyber bullying, CSAM, and other harmful content.

CASE-IN-POINT

Twitch's AutoMod is a live moderation tool that uses NLP and ML to block risky messages across four categories: discrimination, sexual content, hostility, and profanity.

~90% of the videos taken down by streaming platform YouTube are removed by AI-powered automated flagging systems.[3]

However, they need to overcome inherent challenges to be completely mature solutions

Certain aspects of technology warrant careful consideration and improvement to fully harness their future potential. Current challenges include:

- Lack of contextual understanding in more nuanced cases
- Biases due to lack of data, outdated algorithms, or inherent bias of human developers
- Limited understanding of niche languages and cultures
- Challenges in keeping pace with technologies such as GAI
- Lack of transparency in collection and storage of user information
- Privacy and security issues, including data leakage and fraud

By addressing these concerns, enterprises can unlock the future potential of AI technologies and deploy them at scale to optimize their T&S operations. Examples include improved NLP to detect underlying tone and intent of speech, multiple language capabilities to facilitate niche language moderation, and multi-channel behavioral pattern recognition to augment universal user profiling. Enterprises can further train these technologies to understand localized context to formulate localized policies for regulatory compliance.

Emerging technologies are reshaping the digital landscape

Emerging platforms and technologies such as metaverse and GAI have the potential to completely redefine user experience in the digital world. Whether it's the promise of virtual social interactions that can mimic the physical world or the ability to use Large Language Models' (LLMs) multimodal search capabilities, they have the capability to alter the digital landscape forever.

The **metaverse** is a virtual world built on Web3 technology that facilitates social connection and highly immersive user experiences through Augmented Reality (AR) and Virtual Reality (VR). It offers enterprises and users numerous opportunities, such as investing in and trading digitalized assets – Non-fungible Tokens (NFTs), cryptocurrency, etc. – immersive marketing, superior customer experience, far-reaching market access, enhanced employees training, and live virtual events.

However, this virtual world has significant T&S challenges such as abuse of virtual avatars, financial fraud in NFT and cryptocurrency transactions, data privacy concerns, and ambiguous moderation regulations, among others. (For more information on this topic, see our report, Taming the Hydra: Trust and Safety (T&S) in the Metaverse.)

3 Google Transparency Report

Generative AI (GAI) is a form of AI that leverages existing datasets to create new data and text, audio, and visual content. This technological breakthrough shows the potential to transcend the boundaries of creativity as we know it. While individual users employ GAI platforms to solve general queries and generate content, various industries – ranging from multi-format content synthesis for marketing and entertainment to drug discovery for healthcare and life sciences – are exploring its opportunities and use cases.

GAI platform ChatGPT hosted an average of **13 million** users per day in January 2023.[4]

Although GAI has made an enormous contribution to emerging technological advances, it also poses valid concerns around overwhelming spread of unacceptable content, including:

- Copyright infringement and plagiarism
- Veracity of content
- Inherent bias in the underlying algorithms

Furthermore, GAI can be used to mimic human language and generate complex computer codes, giving hackers and criminals unrestricted access and potential to scam users through various cyber attacks.

While GAI and the metaverse currently pose significant concerns on their own, their interplay in the future will give rise to an array of challenges such as regulation of GAI-generated virtual identities in the metaverse, moderation of GAI content in the metaverse, and ownership of GAI-generated 3D assets. The ambiguity of regulations around both these technologies will exacerbate the complexity of deploying GAI in the metaverse.

However, GAI has enormous potential to be the moderation engine of the future, given its capabilities to handle large volumes of data at speed and increasing contextual accuracy. Some potential use cases of GAI in T&S include:

- Generation of synthetic data to improve training models for content moderation
- Improved content safety through the revision of harmful content elements
- Assistance with policy formulation and advice
- Agile and scalable solutions to handle content diversity and volume

To harness the true potential of these AI solutions, the industry is investing to bring them together onto a single platform. Additionally, organizations are working to make sure that such AI solutions are responsible and sustainable for widespread adoption.

4 Similarweb

The ideal trust and safety AI intervention is a holistic platform-based solution

Everest Group take

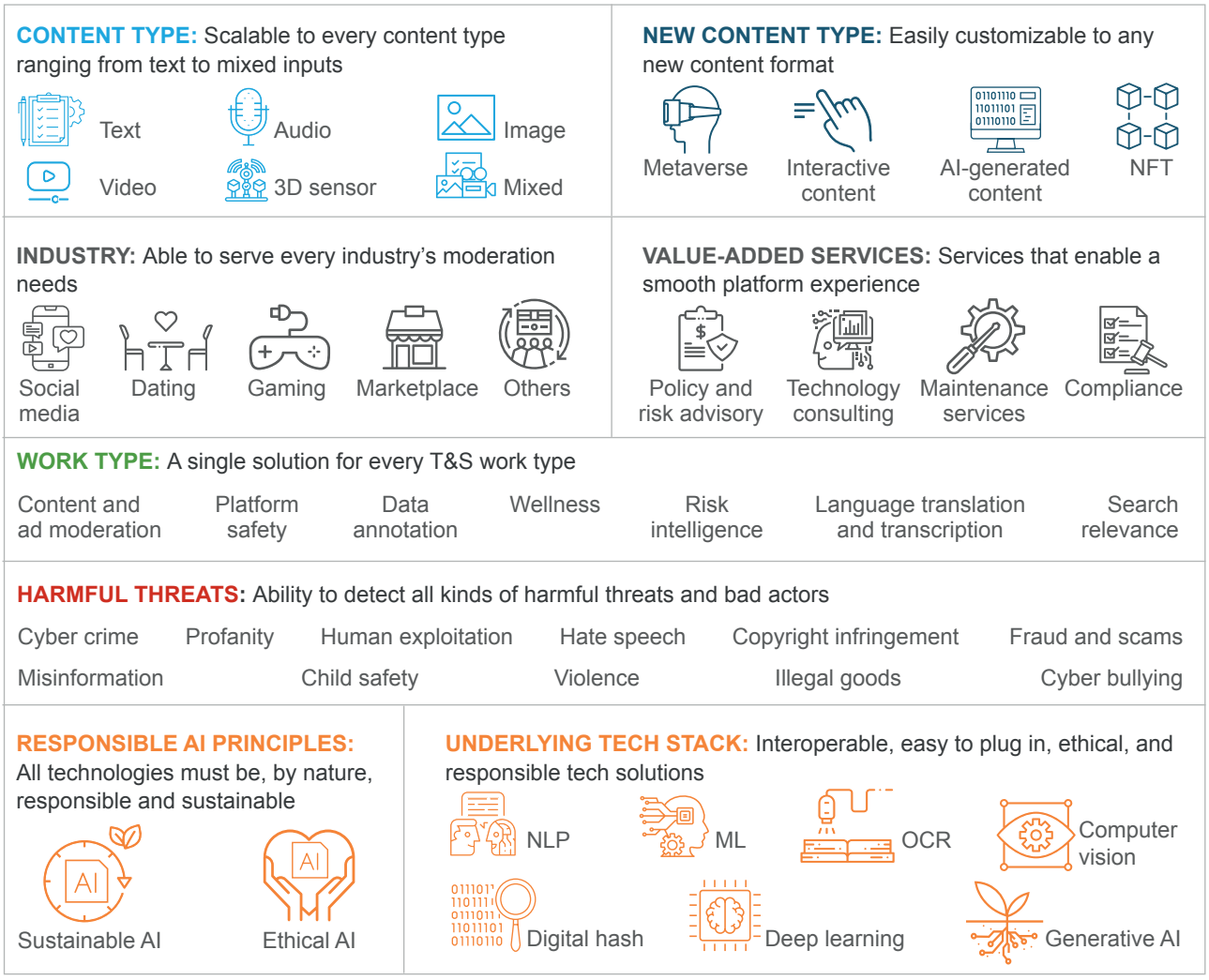
The ideal T&S solution is a holistic, use-case agnostic, platform-based solution that integrates standalone technologies and adheres to the principles of responsible and sustainable AI to ensure a transparent and ethical approach to digital safety.

The current T&S technological landscape is widely fragmented, with multiple AI solutions existing for different uses. A holistic platform can bring all these technologies together to synergize them and serve T&S use cases across industries, work types, and content types, as Exhibit 3 shows.

EXHIBIT 3

The ideal enterprise AI intervention is a holistic, use case-agnostic, platform-based approach

Source: Everest Group (2023)



This integration of such technologies into a single platform will:

- **Simplify moderation** of new content formats and threats with customized moderation queues and codeless workflows
- **Reduce redundancies** in workflows
- **Eliminate moderator need** to navigate different technology interfaces
- **Offer flexibility** for enterprises to model the platform specific to industry and software needs
- **Allow advanced data analytics** by combining data from multiple sources to help improve user profiling and policy design
- **Enable standardized moderation** by facilitating a consistent policy operated moderating procedure
- **Boost solution agility and scalability** with the evolving market
- **Ensure coverage across industries** such as social media, gaming, and e-commerce
- **Facilitate moderation of content created using proprietary LLMs** by combining user preferences, preventing copyright violations, and eliminating biases in training data sets

An all-encompassing T&S platform should combine technologies and incorporate innovations to further increase contextual understanding of content beyond specific keywords through intent and pattern recognition. It will help enhance user experience by removing biases, and proactively identify cases of unacceptable content such as bullying and CSAM disguised under seemingly harmless conversations.

GAI holds the transformative power to be the cornerstone technology in T&S platforms

GAI uses multimodal capabilities to generate both graphic and textual content. Combined with deep learning, it can focus on user prompts and feedback on malicious content and generate more synthetic content to train the platform.

While all AI models have self-learning capabilities, with its reinforcement learning capabilities, GAI can extract insights on previous iterations of content moderation such as harmful content the platform has missed, user dissatisfaction in certain instances, and incorrectly flagged content. It can further use these insights to generate content for additional training models and for policy and risk advisory services. It can also be used as a real-time query resolution system for moderators who want quick access to additional information on platform policy, local regulations, etc.

All technologies will, by nature, have to be responsible and sustainable to ensure long-term growth

With its unprecedented capabilities and immense potential, AI has been a game changer for trust and safety. However, a primary concern around the deployment of AI-powered tools for content moderation is the lack of transparency in its decision-making algorithms. Users have raised concerns around AI models including threat to privacy and discrimination. Consequently, responsible AI is needed to ensure accountability for AI's decisions and actions.

Responsible AI refers to development and deployment of AI that ensures fairness and transparency in operations and safeguards all stakeholders including the platform, users, and regulators. Some of the key elements of responsible AI principles include:

- Preventing AI bias by using diverse and inclusive training datasets
- Using deep learning to make AI models more contextual and nuanced in their decision making

- Providing transparency on automated content-moderation mechanisms
- Ensuring AI's alignment with enterprises' ethical values about individual rights, privacy, and safety

As with any other technology, the development of an AI T&S platform should also consider long-term impact on ecological wellbeing. Training and development of AI models consumes huge amounts of energy and causes excessive greenhouse gas emissions. Enterprises need to be cognizant of the environmental impact of their T&S AI platforms and incorporate sustainable AI principles to limit their negative environmental impact.

Sustainable AI refers to the creation and implementation of AI systems in a way that aligns with sustainable business practices.

Adoption of sustainable AI entails:

- Selecting energy-friendly ML architecture and training models
- Confirming the relevance and quality of data collected and used and removing redundancies
- Optimizing AI hardware to make the infrastructure more energy efficient
- Storing data in more carbon-friendly regions that use alternate forms of electricity such as hydroelectricity
- Considering the life cycles of AI models and responsibly recycling / disposing of obsolete hardware

Integration of responsible and sustainable AI principles into the T&S platform will ensure platform accountability and transparency in user data handling by AI and reduce inherent biases in algorithms. These principles will aid enterprises in aligning and complying with their overall Environment, Social and Governance (ESG) policies.

Training some AI models can emit up to 626,000 pounds of CO2 – equivalent to five times the lifetime emissions of an average American car.[5]

5 MIT Technology Review

Amplify FOCUS on trust and safety to maximize returns on AI investments

Everest Group take

Taking a use-case based approach to technology adoption is key to making rational investment decisions. To leverage the full scale potential of technologies, enterprises need to collaborate with service and technology providers, regulators, and other industry stakeholders.

While early adoption of emerging technologies is important due to diminishing life cycles, it is more important to be right than be first. Enterprises should adopt a use case based approach to proactively evaluate the pros and cons of emerging technologies before charting their investment strategies. There are several considerations that enterprises need to keep in mind while adopting the T&S AI strategies:

- **Identification of AI use cases:** In addition to moderating content, AI-based systems have multiple trust and safety use cases, such as guiding content moderators in decision making, identifying gaps in safety infrastructure, and demand forecasting. Enterprises need to identify such use cases and invest in the right systems to build safe digital platforms for users.
- **Challenges associated with implementation of AI:** Implementation of AI-based systems into the platform can present a variety of complexities such as training the models with high-quality and diverse data, ethical and regulatory concerns, integration of new technologies with the existing infrastructure and impact of such technologies on the wellbeing of moderators.
- **The role of humans:** AI is not mature enough to take over the jobs of human moderators. Human intervention needs to complement technology capabilities. With technological advances and platform integration, moderator role and skillset requirements will evolve. Enterprises need to train and upskill moderators to understand and work with AI technologies.
- **Regulatory compliance:** While a diverse content moderation regulatory landscape is already impacting enterprise policies significantly, AI regulations will add another layer of complexity.

CASE-IN-POINT:

The European Union's proposed Artificial Intelligence Act intends to foster a balance between innovations in AI and protection of the public from its potential harms by strengthening regulations on data collection, transparency, bias, human oversight, and overall ethical accountability of such technologies. To achieve this, the act proposes to classify technologies into four risk tiers – unacceptable, high, limited, and minimal – based on their impact on an individual's overall safety and fundamental rights.

Everest Group's FOCUS framework helps organizations to streamline their platform strategies by bringing attention to five key areas, as described in Exhibit 4.

EXHIBIT 4

Everest Group proposes the FOCUS framework to streamline the enterprise AI strategy

Source: Everest Group (2023)

Foster human-AI synergies	Observe and update the T&S platform regularly	Comply with AI regulations	Upskill content moderators and employees	Strengthen the collaboration ecosystem
Leverage AI for the first level of screening to flag harmful content	Technology updates will make the platform nimble and industry agnostic	Align with local content moderation, data privacy, and especially emerging AI regulations to avoid penalties	Retrain moderators through training and workshops to leverage the AI platform	Partner with service and technology providers to acquire and update the latest technology solutions
Human moderation can offer contextualization, fairness, and flexibility	Cybersecurity and data privacy updates will prevent exposure to ongoing threats	Conduct regular audits and maintain a transparent moderation policy	Use AI tools such as conversational AI for moderator queries and wellbeing	Brainstorm with researchers, policymakers, and regulators to address ethical moderation and policy concerns

Collaborations with third-party providers can enhance the enterprise AI investment strategy

Enterprises have generally developed T&S technology capabilities largely in-house. However, collaboration with service and technology providers can augment enterprise AI investment strategies in several ways:

- **Provide policy management support:** Providers assist with constant market monitoring in addition to collaborating with regulatory bodies and policymakers to identify triggers that can potentially impact existing T&S policies.
- **Moderator wellbeing:** Providers leverage various existing and emerging technologies to deliver cutting-edge wellbeing solutions for content moderators. Partnerships with these providers can help expand the scope of enterprise wellbeing interventions.
- **Deliver holistic platform-based T&S solutions:** Service providers are increasingly partnering with technology providers that offer comprehensive, scalable platforms integrating various content handling technologies and human capabilities into unified systems.
- **Approach innovation collaboratively:** The right partner can help explore new technological innovations, such as explainable AI and responsible and ethical AI, to detect patterns and anomalies and assist enterprises in keeping their platforms secure.

- **Customize moderation rules:** Providers can assist enterprises in implementing customized moderation rules that are specifically geared to identify and highlight outputs violating enterprise guidelines or other policies.
- **Upgrade algorithms and training datasets:** Partnerships can help maintain enterprise T&S platform robustness by improving algorithm accuracy and including a wide variety of datasets to train the T&S technology.

Enterprises can utilize the tech providers’ technical knowledge and service providers’ T&S expertise to develop scalable AI solution-based T&S platforms that adhere to the enterprises’ AI policies.

Conclusion

AI is a double-edged sword. Rapid evolution of advanced AI technologies will lead to a surge in complex content on the internet. Such content, if unmoderated, can cause significant harm to the users. At the same time, some of these technologies show immense potential in enhancing organizations’ T&S capabilities. Enterprises need to explore the potential of these technologies and scale up investments quickly to keep pace with the technological evolution and the expanding universe of toxic content. Additionally, they need to unify these technologies to develop end-to-end platform capabilities and ensure long-term sustainability in T&S operations.

While focusing on in-house investments is one way to build such capabilities, a collaborative mindset can accelerate the path to a safer internet. A connected ecosystem that comprises enterprises, service providers, and technology providers will drive the evolution of T&S technology. It is essential that enterprises nurture these collaborations by taking the initiative and supporting them with necessary investments. Such initiatives, coupled with knowledge sharing through industry collectives and continued engagement with third parties such as communities, regulators, and universities, can go a long way in achieving organizations’ T&S goals.

Everest Group®
With you on the journey

Everest Group is a leading research firm helping business leaders make confident decisions. We guide clients through today's market challenges and strengthen their strategies by applying contextualized problem-solving to their unique situations. This drives maximized operational and financial performance and transformative experiences. Our deep expertise and tenacious research focused on technology, business processes, and engineering through the lenses of talent, sustainability, and sourcing delivers precise and action-oriented guidance. Find further details and in-depth content at www.everestgrp.com.

This study was funded, in part, by **Wipro**

For more information about Everest Group, please contact:

+1-214-451-3000
info@everestgrp.com

For more information about this topic please contact the author(s):

Manu Aggarwal, Partner
manu.aggarwal@everestgrp.com

Abhijnan Dasgupta, Practice Director
abhijnan.dasgupta@everestgrp.com

Darshita Lohiya, Senior Analyst
darshita.lohiya@everestgrp.com

Aseem Nousher, Senior Analyst
aseem.nousher@everestgrp.com

About Wipro

Wipro has been at the forefront of IT service capabilities for over 4 decades. We have leveraged this to design and enhance our trust and safety solutions for over a decade. We are among the top 10 companies in the world which offers robust trust and safety solutions that cater to various use-cases such as:

- User generated content
- Advertisements
- Social pages
- Communities
- Identifying fraudulent transactions and patterns
- Social media monitoring for brand health and customer sentiment
- Policy advisory

Wipro is committed to building safe digital spaces for our clients and helping their employees bring their best selves to work. We provide comprehensive trust and safety solutions, formulate holistic wellbeing policies, and offer policy consulting and advisory services based on our experience working with some of the world's largest and most diverse organizations.

With over 1.1 billion+ post reviews conducted, our multi-format content moderation capabilities are delivered through a state-of-the-art impact sourcing program and create a balance between human and machine intelligence.

In addition to our capabilities, we prioritize the wellbeing and resilience of our content moderators through wellness programs that ensure their physical and mental health on a daily basis.

Ambitions Realized.

Wipro Limited
Doddakannelli
Sarjapur Road
Bengaluru – 560 035
India

Tel: +91 (80) 2844 0011
Fax: +91 (80) 2844 0256
wipro.com

Wipro Limited (NYSE: WIT, BSE: 507685, NSE: WIPRO) is a leading technology services and consulting company focused on building innovative solutions that address clients' most complex digital transformation needs. Leveraging our holistic portfolio of capabilities in consulting, design, engineering, and operations, we help clients realize their boldest ambitions and build future-ready, sustainable businesses. With 250,000 employees and business partners across more than 60 countries, we deliver on the promise of helping our clients, colleagues, and communities thrive in an ever-changing world.

For more information, please write to us at **info@wipro.com**

IND/iCORE/JUL 2023-JUN 2024