AI's Playdate with Youth Platforms: A Review of AI-Driven Content Moderation Techniques

Aditya Surve[1*†], Raghav Jain[2*†], Mokshit Surana[1*†],
Sainath Reddy Sankepally[3†], Gautam Malpani[4†],
Swapneel Mehta[5*]

[1*]Information Technology, Thadomal Shahani Engineering College, Mumbai, Maharashtra, India.
[2*]Indian Institute of Technology, Patna, India.
[4]Information Technology, Dwarkadas Jivanlal Sanghvi College Of Engineering, Mumbai, Maharashtra, India.
[3]Artificial Intelligence, IIIT Raipur, India.
[5]Questrom School of Business, Boston University, USA.

*Corresponding author(s). E-mail(s): swapneel@bu.edu;
†These authors contributed equally to this work.

Abstract

The exponential rise of user-generated content across digital platforms has rendered manual content moderation inadequate, making AI integral to this process. Through natural language processing and computer vision technologies, AI can rapidly scan massive volumes of data to flag policy violations with unparalleled speed and scale. In practice AI-based approaches are imperfect, suffer from biases, and cannot be end-to-end automated due to the high cost of false positives. In light of the challenges facing content moderation, this literature review examines the role of AI techniques for content moderation across three key areas with a high volume of younger users – social media, dating apps, and gig economy platforms. Publicly available materials provide insights into the moderation approaches adopted by leading youth-centric social platforms like Snapchat, Tinder, TikTok, and Twitch. This review discusses the use of AI-based techniques for hate speech detection, age verification, misinformation identification, and others. We focus on identifying the common problems prevalent across individual platforms belonging to each industry, in an effort to provide a common ground for the discussion of shared solutions in these spaces. We find that while AI enables more nuanced and comprehensive moderation across the industries, there are limitations with respect to contextual understanding, bias, and adversarial generation of synthetic content. Continuous improvements to training data diversity, model transparency, and human oversight remain vital to advancement. By harnessing AI as a scalable first line of defense, platforms can uphold standards of safety, authenticity and dignity for their users. On the other hand, sole reliance on algorithmic systems is unwise given their known and unknown limitations. Blending automated flagging with human content reviewers appears the most prudent moderation strategy but generative AI technologies are likely to exacerbate the challenges of content moderation.

1 Introduction

In the span of just a few decades, the digital world has witnessed the meteoric rise of social platforms, evolving from nascent online communities into digital behemoths. Often conceptualized as spaces for local communities to connect, these platforms have burgeoned into global networks; shaping cultures, influencing trends, and molding societal narratives. For the youth, in particular, these platforms are more than just tools to facilitate connection; they are integral to identity formation, socialization, learning, and activism. Whether it's rallying for social causes on Twitter, expressing creativity on TikTok, or building communities on Instagram, young individuals have leveraged these platforms to amplify their voices, share their stories, and connect across boundaries. As these platforms have become intertwined with their everyday lives, the question arises: how do these platforms prioritize user safety and content integrity for their audiences? To answer this question, we identify public materials on social platforms that cater to youth, and summarize the challenges and opportunities for online safety presented by the use of artificial intelligence (AI) and machine learning (ML) technology, including the impact of generative AI (GenAI) models. We dive into the intricacies of content moderation through these resources, verifying our findings through interviews with four anonymous industry sources that have recently been involved in leading such work at youth social platforms. We identify candidates that qualify within the industries of Social Networks, Online Dating, or Gig Economy. We find these three areas to be emblematic examples of a mature industry, a fast-growing industry, and an emergent industry, respectively, with each candidate within these industries demonstrating a marked presence of youth as users or providers on their platforms.

As the digital landscape evolves, the emergence and maturation of social network platforms like Snapchat, TikTok, and Instagram have redefined how the younger generation interacts, communicates, and expresses themselves. From serving as platforms for virtual connections, they have solidified their positions as influential hubs of digital communication. The dynamic nature of Snapchat's ephemeral content, TikTok's short-form videos, and Instagram's visual narratives has propelled them beyond mere entertainment outlets. Instead, they have become integral to the fabric of young individuals' lives, catalyzing the development of personal identities, fostering social bonds, and serving as conduits for creative expression. This evolution prompts an imperative inquiry into the strategies and mechanisms that govern these mature social platforms, ensuring their continued alignment with user needs, safety, and societal values.

Simultaneously, the digital landscape is witnessing a spurt of growth in another distinctive sector–dating apps. Platforms such as Tinder and Bumble have surfaced as transformative spaces that enable young adults to navigate the complexities of modern relationships and companionship. Driven by the surge in mobile technology and changing social norms–not to mention the COVID-19 pandemic–these apps have swiftly transitioned into mainstream methods of connecting with a partner. The intimate and personal nature of interactions on dating apps necessitates vigilant content moderation strategies to ensure that users can engage in meaningful connections without compromising their safety or emotional well-being. As this growing area garners more attention, the need to comprehend the evolving landscape of content moderation within dating apps becomes ever more pronounced.

Moreover, the ever-evolving digital landscape continues to extend its reach into the realm of the gig economy—an area marked by its rapid emergence and large impact on the younger demographic. Platforms like DoorDash and Uber have introduced an innovative approach to employment, offering individuals short-term, flexible work through online interfaces. This emergent sector has engendered a unique set of content moderation challenges, as platforms are tasked not only with ensuring fair labor practices but also with safeguarding the well-being and dignity of workers. The experiences of gig workers, especially those in the younger age group, are marked by a range of challenges, from rudeness and unsafe conditions to unwanted advances. As this sector takes center stage, setting up effective content moderation within gig economy platforms becomes of paramount importance, necessitating a holistic analysis of strategies and mechanisms to limit online harms.

In the dynamic landscape of these three areas—social networks, dating apps, and the gig economy—the importance of content moderation cannot be overstated. The manner in which content is curated, screened, and managed directly impacts user experiences, safety, and overall platform integrity. Ensuring that the digital spaces inhabited by the youth remain conducive to meaningful interactions, safe exploration, and dignified work is a collective responsibility. Effective content moderation has ripple effects beyond platform operations, influencing the well-being of individuals, societal norms, and the broader digital ecosystem. Therefore, a comprehensive understanding of content moderation strategies in each of these areas is essential to not only foster a secure online environment but also to shape the future trajectory of digital engagement for the younger generation.

With the exponential growth of user-generated content across social media, dating apps, and gig platforms, manual content moderation has become highly inadequate and impractical [1]. This renders AI integral to the moderation process. Through natural language processing and computer vision, AI can rapidly scan massive volumes of textual, visual, and audio data to flag policy violations with unmatched speed and scale [2]. Automated tools powered by machine learning algorithms can detect harmful content like cyberbullying, nudity, violence, scams, hate speech, and misinformation with increasing accuracy [3]. Moreover, AI helps reduce moderator fatigue and emotional toll by handling the bulk of repetitive tasks, while empowering human reviewers to focus on nuanced cases. With continual improvements in AI capabilities, content moderation is growing more effective, nuanced, and comprehensive across diverse digital platforms. Though not infallible, AI enables platforms to better balance free expression with user safety and wellbeing.

The use of AI in content moderation has rapidly evolved over the past two decades. In the early 2000s, basic keyword filtering and blacklist databases marked the first forays into automated moderation. However, these rudimentary tools were limited in scope and understanding. The late 2000s brought more advanced algorithms that could detect prohibited imagery like nudity and violence. Machine learning enabled platforms to train systems to flag hate speech, cyberbullying, and other textual violations by the early 2010s. By mid-decade, AI adoption reached mainstream prominence, as social networks began proactively finding policy breaches before human detection. Deep learning substantively improved accuracy, while natural language processing allowed for more nuanced analysis. Today, state-of-the-art systems like GPT-4 [4] can understand complex linguistic contexts and assess harms more holistically.

While AI has made significant advances in content moderation, it continues to face major failures and limitations. One persistent challenge is contextual understanding - AI still struggles to interpret the nuances and implied meanings in user-generated content [5]. This inability to understand context leads to errors, such as flagging harmless posts as inappropriate. It also means human moderators are still required to monitor nuanced cases that automated systems cannot comprehend [6]. Additionally, AI stumbles with assessing cultural contexts and languages outside of English [7]. Most content moderation systems are trained on English data, rendering them ineffective for other languages. The cultural references and colloquialisms inherent in different languages pose major challenges. Furthermore, cultural contexts regarding what is considered offensive or dangerous varies across geographies, which can confuse AI models. These failures to grasp contextual cues and language/cultural variances underscore that AI cannot fully replace human diligence. While AI provides a useful tool for scaling moderation, its flaws necessitate ongoing human oversight, deliberation and retraining.

The advent of generative AI models like DALL-E and ChatGPT has significantly strained the content moderation ecosystem, as it is now easy and cheaper to rapidly create low-quality content [8]. These models enable easy and rapid generation of textual, visual, audio, and multimodal content, exponentially increasing the volume needing review. Both the quality and quantity of AI-produced content makes comprehensive moderation impractical. Moreover, generative models can smoothly create synthetic media, toxic text [9], misinformation [10], and other harmful content that evades detection. This deluge of AI-produced content has overwhelmed human moderators and made comprehensive oversight near impossible. Moreover, the quality of generative content is approaching human-level creations, making detection harder for both humans and AI systems. The rapidly evolving creativity of generative models enables the swift production of harmful content like deepfakes, text scams, and disinformation that can spread virally before moderation occurs. This delay between generation and moderation increases the potential for abuse and coordinated influence campaigns. Additionally, generative models have enabled easier production of content that toes the line of acceptability and malicious creativity intended to evade filters.

These dynamics spotlight key moderation failures against generative AI's pace and scale.

1.1 Evolution of Social Media Usage Among Youth

Social media usage among youth has undergone remarkable evolution over the past decade. In the early 2010s, engagement centered on entertainment and peer connections. But as platforms matured, habits transformed: the mid-2010s brought identity construction through self-presentation on sites like Instagram, which grew from just 9% of adults in 2012 to 40% by 2021, with 71% of youth ages 18-29 using it [11]. TikTok has now rocketed in popularity among teens, used by 67% compared to just 32% who use Facebook [12]. YouTube leads among 95% of teens, followed by TikTok, Instagram (62%), and Snapchat (59%) [12]. Teen TikTok and Snapchat users are highly engaged, with about a quarter using them almost constantly. Overall, 46% of teens say they are online almost constantly, raising concerns around overuse [12]. Dating apps have also grown among youth, with over half of adults in the USA under 30 having tried them. Tinder leads among 79% of dating app users under 30, facilitating modern dating. [13] The gig economy has also impacted youth, with 30% of adults under 30 having earned money through online gig platforms in the USA [14]. While entertainment remains integral, the platforms now enable identity formation, activism, and community building. The psychological effects of these platforms, addiction and overuse, toxicity, poor safeguards, and lack of protections are growing concerns. While advanced economies are observing a plateau in internet penetration and social media use, developing countries are witnessing a surge. This expansion of digital platforms in the developing world has paved the way for a more interconnected global society. From 2013-14 to 2017, internet usage in emerging economies escalated from 42% to 64%. In contrast, advanced economies have been relatively stable with around 87% internet penetration [15]. Developing economies are rapidly adopting social media platforms. In 2015-16, roughly 40% of adults in these countries were on social networking sites. This number climbed to 53% by 2017. In contrast, developed countries observed a relatively stagnant trend in social media use [15].

1.2 The Impact of COVID-19 on Online Content Consumption

The COVID-19 pandemic led to major changes in how people consumed content online. With lockdowns and social distancing measures in place limiting physical movement severely, most of us turned to the internet for information, entertainment, social connection, shopping, and other daily needs. Video calling became ubiquitous, with 81% of Americans reporting using video calls during the pandemic. Platforms like Zoom and FaceTime became vital for work meetings, school, social gatherings, and maintaining relationships when in-person contact was not possible. However, 40% of frequent video call users reported feeling fatigued or worn out from excessive video call time. Beyond video, internet usage surged more broadly. 90% of Americans said the internet was essential or important during the pandemic. 40% reported using technology and the internet in new or different ways compared to pre-pandemic. From online grocery shopping to telehealth visits, many activities migrated online that had traditionally

been done in-person. With schools closed, remote learning highlighted the "homework gap" wherein lower income students lacked proper devices or internet access to complete assignments at home. This led to greater public support for schools providing laptops/tablets to students, with 49% now supporting this policy compared to just 37% in April 2020. Media consumption also increased, with 72% of parents reporting their children were spending more time on screens during the pandemic [1].
Developing countries including India saw a similar shift in internet usage arising from the COVID-19 pandemic. A 2021 study [2] found 61% of Indian households used the internet, up from just 21% in 2017. The 15-65 age group saw internet use rise from 19% to 49% during the pandemic. An estimated 80 million Indians came online in 2020, with 34 million doing so specifically due to COVID-19 impacts. This shift of a large number of users to a digital-first mode of operation was accompanied by changes to user mobility, lifestyle, and created novel modes of interaction in the digital realm. It also meant that a large number of newly minted digital platform users became vulnerable to financial fraud, scams, and other online harms perpetrated through the abuse of the very same social platforms.

2 Related Work

The practice of content moderation across various digital platforms has witnessed a remarkable transformation. Social platforms have emerged as pivotal tools for information dissemination, personal expression, and entertainment. With this profound influence comes an equally significant responsibility to manage and monitor the content that users share and access. However, as these platforms mushroomed in popularity, they faced a barrage of challenges concerning the content users were generating and disseminating. Some notable challenges included the prevalence of hate speech, harassment, misinformation, and cyberbullying.

2.1 Challenges by Industry

1. **Social Media Platforms:** The digital terrain was primarily disrupted by the appearance of fake news, hate speech, and cyberbullying.
2. **Video-based Platforms:** Platforms hosting video content grappled with copyrighted material, hate speech, and explicit content issues. Live streaming further complicated these issues, with platform policy violations arising during real-time broadcasts, with difficulties in proactively limiting harassment and inappropriate content.
3. **Instant Messaging Services:** IM services had to address concerns like spam, phishing, inappropriate content sharing, and privacy breaches of user data.
4. **Global South Social Platforms:** With expansion into multilingual societies that actively shared multimodal content, platforms faced severe data limitations leading to inefficient automated classifiers. Furthermore, the complex demographic and cultural contexts in these regions led to challenges in defining uniform platform policies.

[1] https://www.pewresearch.org/internet/2021/09/01/the-internet-and-the-pandemic/
[2] https://www.indiatimes.com/technology/news/india-internet-usage-report-554181.html

2.2 Platform Policy Initiatives

There have been various platform policy initiatives notably the release of so-called Transparency Reports that are designed to yield information into the trust and safety and content moderation operations of platforms. We list examples that are selected to highlight a variety of platform-specific challenges across a broad spectrum of social media platforms and modalities of online harms:

1. **Meta (formerly Facebook):** In 2020, Meta unveiled a whitepaper addressing pertinent issues like defining what constitutes ”harmful content,” accountability of online platforms, effective content regulation, and possible performance targets [16].
2. **Snapchat:** Their Transparency Report for the latter half of 2022 showcased a significant increase in content and account reports, particularly in the Harassment Bullying and Other Regulated Goods categories [17] [18]. Furthermore, Snapchat addressed legal requests from law enforcement and governments, evidencing the broader societal implications of content moderation.
3. **YouTube:** Their whitepaper highlighted the immense scale of content regulation, with YouTube removing millions of videos for various guideline violations, ranging from harmful behavior to impersonation [19] [20] [21].
4. **Twitter:** Their Social Media Compliance Reports showcased Twitter's proactive steps in suspending accounts involved in child exploitation and terrorism [22].
5. **WhatsApp:** They witnessed fluctuations in account bans across several months, highlighting the dynamic nature of content moderation challenges [22].
6. **Tinder:** Addressing concerns of harassment and inappropriate interactions, Tinder has integrated automated screening for offensive messages. However, the volume of such messages hasn't decreased, indicating the persistent challenges platforms face [23].
7. **TikTok:** Their Community Guidelines Enforcement Reports illustrated the enormity of content regulation, with millions of videos removed for various violations [24].

Content moderation remains at the forefront of challenges faced by social platforms. The ongoing efforts by major platforms showcase the importance of continuous adaptation, technological innovations, and robust policies to ensure user safety and maintain the integrity of online communities.

3 The Impact of Generative AI on Content Creation

Artificial Intelligence, especially Generative AI (GenAI), has drastically altered the landscape of content creation, democratizing the content generation process and allowing virtually anyone to become a creator participating in the digital economy. However, as with most technological advances, the implications of this transformation are both promising and potentially problematic.

3.1 Democratization of Content Creation

1. **Accessibility for All:** With tools powered by GenAI, individuals no longer need to be experts to generate content. From producing music tracks to writing articles, AI offers templates and frameworks that make the process accessible to novices.
2. **Personalization:** GenAI can tailor content to individual preferences. This capability can be particularly useful for businesses looking to create personalized marketing campaigns or educators developing customized learning materials.
3. **Efficiency and Scale:** Creating content, especially in bulk, can be time-consuming. GenAI tools can expedite this process, generating large volumes of content in a fraction of the time a human would take.

3.2 Challenges and Concerns

1. **Authenticity Issues:** With the ease of content generation comes the challenge of authenticity. It becomes harder to discern if a piece of content is original or machine-generated, leading to potential trust issues. There have been efforts to identify content created by artificial intelligence, including AI text classifiers [25] and watermarking techniques [26]. However, these approaches have limitations in their accuracy and reliability [27], so they should not be the sole basis for making real-world decisions. More research is needed to develop robust methods for detecting synthetic content generated by AI systems.
2. **Over-saturation:** The digital space is already crowded with content, and GenAI might exacerbate this problem. With everyone being given the ability to produce content for cheap, platforms may become over-saturated with low-quality content. Among other issues, this also makes for a worse user experience since high-quality content may receive lower exposure on average.
3. **Misuse and Disinformation:** The ability to generate content rapidly and in large volumes can be misused. There is potential for spreading misinformation, fake news, or propaganda, especially if malicious actors harness the capabilities of GenAI.

4 AI-based Content Moderation Techniques

The explosive increase in user-created content on digital platforms has surfaced many troubling issues, including hate speech, misinformation, adult content, and other dangerous narratives. Protecting online communities without suppressing authentic discussion poses a formidable challenge. AI has emerged as a potentially game-changing solution for content moderation in this landscape.

4.1 Broad-Spectrum Detection Capabilities

AI, leveraging sophisticated algorithms, can span a wide spectrum of inappropriate content:

1. **Hate Speech and Toxicity:** Advanced Natural Language Processing (NLP) algorithms can discern language patterns associated with hate, discrimination, or general toxicity, ensuring that such content is flagged or filtered out [28–31].
2. **NSFW Content:** AI models trained on vast image and video datasets can effectively identify explicit, graphic, or otherwise Not Safe For Work (NSFW) content, preventing its dissemination on platforms where it's inappropriate [32, 33].
3. **Misinformation and Fake News:** AI can analyze vast amounts of data rapidly, comparing new content against verified databases, to flag potential misinformation or false narratives [34, 35].

4.2 Proactive Content Moderation

Instead of post-hoc solutions after inappropriate content is posted, AI allows for a proactive approach:

1. **Real-time Feedback:** As users compose messages or upload content, AI can provide instantaneous feedback on potentially problematic content, fostering self-awareness and self-moderation among users. Counter-narrative response generation [36] and intervention strategies [37] have been developed by researchers attempting to tackle this problem.
2. **Content Neutralization:** AI does not just flag inappropriate content; advanced models can also suggest or implement modifications to make the content compliant, ensuring that the core message remains intact without the harmful elements. Researchers have developed detoxification techniques to neutralize toxic and offensive content [38, 39].

4.3 Industry Adoption of AI for Content Moderation

Major social media platforms are turning to AI to enforce their content policies and keep their communities safe. Meta (Facebook) uses machine learning models to proactively detect and remove policy-violating text, images and videos at scale [40, 41]. Similarly, Snapchat and Twitter leverage AI to moderate content based on their guidelines [42]. Dating apps like Tinder and Bumble also employ AI to detect offensive language and harassment, improving safety for their users [43]. Beyond social media, AI moderation is being adopted across diverse digital platforms including e-commerce sites, forums, blogs and gaming platforms. Airbnb, Twitch, Reddit and others use AI to flag policy breaches in user-generated content. In recent years, numerous startups have emerged focusing specifically on AI-driven content moderation services. Upcoming startups like Cove [3], and Sero AI [4] aim to offer APIs and tools to automate content screening for online platforms that are designed in accordance with platform use-cases.

[3] https://blog.getcove.com/safety-layer

[4] https://www.getsero.ai/

Category	**Apps**	**Fake Profile**	**Online Harassment**	**Hate Speech**	**Underage Users**
Dating Apps	*Tinder*	facilitate catfishing, scams, destroying user trust	breeds toxicity, conflicts with secure matches	ostracizes minority users	risk grooming, exploitation
	Bumble	undermine user identity credibility	jeopardizes safety, violates respect commitment	breeds toxicity, fear	Same as Above
	OkCupid	sabotage algorithms reliant on truth	Same as Above	obstructs expression, marginalizes users	Same as Above
Gig and Streaming Platforms	*Uber*	fraudulent activities, erodes trust	corrodes user trust, safety	enables discrimination	risks of assault, harm
	Twitch	view botting, inflated counts	cultivates toxicity	cultivates discrimination	violence, grooming risks
	Doordash	fraudulent refunds, account theft	jeopardizes safety	erodes service reliability	violates labor laws
Social Media	*Snap*	impersonation, fraudulent communication	impedes expression, damages connections	damages cohesion	inappropriate content exposure
	Instagram	inauthentic influencer growth	cultivates toxicity, inhibits expression	cultivates discrimination	Same as Above
	Sharechat	spread of misinformation	silences voices, spread of hate	endangers minority users	radicalization risks
	TikTok	artificial inflation of trends	hampers expression, damages wellbeing	cultivates discrimination	predators, mature themes risks

Table 1 An Overview of Traditional Integrity Concerns Across Digital Platforms: This table provides a comprehensive breakdown of traditional integrity challenges associated with various digital platforms, categorized under themes such as Fake Profile, Online Harassment, Hate Speech, and Underage Users. It distinctly highlights how different platforms, from dating apps to social media, grapple with these concerns and how users might be affected. The data underscores the importance of robust moderation and preventive measures to ensure user safety and trust.

Category	**App**	**AI-generated Media**	**Fake Personas (Interactive)**	**NA**
Dating Apps	**Tinder**	facilitates more convincing fake profiles and catfishing	Interactive fake personas allow more manipulative catfishing	NA
	Bumble	Same as Above	Same as Above	NA
	OkCupid	Same as Above	Same as Above	NA
Category	**App**	**Hallucinations and Biases for Chatbots**	**Impersonation**	**AI-enabled Fraud**
Gig and Streaming Platforms	**Uber**	can enable discrimination against riders	facilitates account theft	AI synthesis of fake photos and ride receipts
	Twitch	can spread toxicity and misinformation	facilitates scams and copyright infringement	AI chatbots, deepfakes, and view bots
	Doordash	enable discrimination against delivery workers	Same as Uber	AI-generated phishing links and fake profiles
Category	**App**	**AI-generated Media**	**Bots**	**Influence Ops**
Social Media	**Snap**	AI-altered media can spread misinformation	disseminate deceptive snaps	orchestrate snap stories to push agendas
	Instagram	AI-enhanced photos may set unrealistic standards	misinformation through auto-posted content	curate and promote divisive narratives
	Sharechat	Same as SnapChat	Same as Above	Orchestrated campaigns can manipulate users
	TikTok	AI-generated clips can propagate misleading content	promote videos containing misinformation	engineer viral video trends

Table 2 Insights into AI-related Integrity Concerns: The table elucidates a wide range of integrity challenges that arise from the integration of artificial intelligence technologies across different digital platforms. It categorizes these challenges based on different themes such as AI-generated media, impersonation, and AI-driven fraud, providing a detailed look into how platforms like dating apps, gig and streaming platforms, and social media handle and could potentially be misused with AI capabilities.

Category	**App**	**Romance Scams / Catfishing**	**Personal Safety**	**Cyberbullying**
Dating Apps	**Tinder**	can involve users presenting fake personas, exploiting emotions for financial gains or personal leverage	Inadequate screening and trackability around real world meet-ups jeopardize personal safety,	creates toxic experiences that dissuade usage
	Bumble	Same As Above	Same As Above	Same As Above
	OkCupid	Same As Above	Same As Above	Same As Above
Category	**App**	**Hallucinations and Biases for Chatbots**	**Personal Safety**	**NA**
Gig and Streaming Platforms	**Uber**	Fake ride requests and mileage inflation scams	Inadequate driver background checks and ride monitoring jeopardize passenger safety	NA
	Twitch	Donation scams and fraudulent subscriber gifting	Minimal oversight around stream meetups and conventions jeopardizes safety	NA
	Doordash	Fake deliveries, fraudulent refunds, and account takeovers	Lack of screening around delivery worker identities and activities endangers personal safety of customers and merchants	NA
Category	**App**	**Mis/Disinformation**	**Doxxing**	**Parental Monitoring Problem**
Social Media	**Snap**	Disinformation campaigns on Snapchat, leveraging ephemeral yet viral content, compromise truthful public discourse	severely undermine user privacy and freedom of expression	The disappearing nature of Snapchat content inhibits parental monitoring
	Instagram	Misinformation and fake news on Instagram, amplified by bots and inauthentic likes, spreads ignorance,	Doxxing and privacy breaches on Instagram damage user safety and self-expression	Instagram's expansive content and opaque algorithmic feed make parental monitoring difficult
	Sharechat	Misinformation in regional languages, amplified by bots, threatens user safety and inclusion	jeopardize vulnerable users and marginalized groups,	Lack of parental controls in ShareChat's regional languages threatens child safety
	TikTok	Highly-shareable misinformation on TikTok distorts public discourse	violate user safety and law, conflicting with its stated commitment to community protection	TikTok's addictive infinite feed algorithm and privacy features obstruct parental monitoring, risking overuse and inappropriate content access.

Table 3 Overview of Platform-specific Integrity Concerns: This table provides a comprehensive breakdown of potential issues across popular youth Social platforms, categorizing concerns into different areas such as romance scams, personal safety, and cyberbullying. It offers insights into the multifaceted challenges platforms face in ensuring user safety and trust.

5 Challenges and Opportunities for Content Moderation

5.1 Social Media Apps

Social media platforms have transformed from mere channels of communication into integral facets of our daily lives, shaping our interactions, opinions, and our perception of the world. Even as their reach expands and their influence deepens, they are grappling with a host of complex challenges. Specifically, in this industry, we examine four prominent social media applications ('apps') - Snapchat, ShareChat, TikTok, and Instagram - each with its unique user base and affordances. The intricacies of the issues faced by these platforms not only test the ethical and operational capabilities of these tech giants but also raise critical questions about the responsibilities of digital platforms in society. This paper lays out some of the core challenges faced by these platforms, offering evidence of their prevalence and shedding light on their broader implications for both users and the platforms. Our aim is to highlight common challenges and solutions for each industry, across each of these platforms, in order to help the public and platforms identify areas where open collaboration might improve online safety, which has been a contentious topic in light of regulatory discussions.[5]

1. **SnapChat**
   (a) Ephemeral Content and Parental Monitoring Problem: The disappearing nature of content, which vanishes after being viewed, poses significant challenges for parental monitoring[47].
   (b) Location Sharing Concerns Problem: Snap Map's feature, which shares users' real-time locations, raises safety and privacy concerns. Reports from The Guardian [48] and other sources [49] highlight the potential risks associated with location sharing on Snap Map.
2. **Instagram**
   (a) Volume and Content Moderation Problem: The massive amount of daily content makes manual review cumbersome and at times, impractical. With over 2 billion monthly active users, as per data from Hootsuite [50], Instagram faces significant challenges in effectively moderating content.
   (b) Mental Health of Content Moderators Problem: The exposure to graphic and distressing content takes a heavy psychological toll on human content moderators. "The Underworld of Online Content Moderation" from The New Yorker illuminates the traumatic experiences of moderators [51].
   (c) Deepfakes and Misinformation Problem: The rapid proliferation of deepfakes challenges content moderation systems. An analysis from Brookings Institution underscores the difficulties in identifying and managing deepfakes on platforms like Instagram [52].
3. **ShareChat**

[5] TikTok has a staggering 1.1 billion monthly active users worldwide and over 220 million downloads in the U.S. alone. Particularly resonating with the youth, 60 percent of its U.S. audience falls within the 16-24 age bracket, spending an average of 95 minutes per day on the app. TikTok's impact is further solidified by its 500 million dollars in U.S. revenue in 2020, a thriving influencer marketing ecosystem, and a competitive advertising landscape [44]. Meanwhile, Instagram, boasting 1.6 billion monthly active users as of April 2023.It outpaced Facebook with a 12.2 percent surge, adding 176 million users in the past year. Instagram's diverse user base, led by India with 326.6 million users, is predominantly composed of 18-24 year olds (32 percent) and 25-34 year olds (29.6 percent) [45]. On a different note, Snapchat secures its position as the 10th most popular social media platform globally with 750 million daily active users. Catering to a younger demographic, Snapchat's user base is concentrated between 15-25 years old, reaching over 90 percent of 13-24 year olds in more than 20 countries. With India leading with 182 million users, Snapchat's engaging features, including augmented reality filters utilized by over 250 million daily active Snapchatters, contribute to its substantial presence [46].

   (a) Hate Speech and Vulnerable Communities: The platform is vulnerable to issues of hate speech, especially against marginalized communities. Reports from Hindustan Times shed light on the rampant misinformation and hate speech on regional platforms like ShareChat [53].
   (b) Language-based Moderation Problem: Moderating content in various regional languages adds layers of complexity as sharechat is one of the most popular app in regional areas in India. With over 15 Indian languages supported, as noted in TechRadar, ShareChat grapples with the intricacies of moderating content in multiple languages [54].

4. **Tiktok**

   (a) Viral Challenges and Safety Concerns: The platform's viral nature sometimes promotes challenges that pose risks of users engaging in dangerous activities. Numerous reports, including those on platforms, detail various risky challenges that have gone viral on TikTok [55, 56].
   (b) Copyrighted Material Usage Problem: Unauthorized use of copyrighted music or visuals is prevalent. Insights from Digital Information World highlight the challenges TikTok faces with copyright violations [57].
   (c) Scam Promotions Targeting Young Users Problem: Young users on the platform are frequently targeted by scam promotions. Reports from The Guardian elucidate how the platform's algorithm may inadvertently promote scams [58].

5.1.1 Safety measures taken by these platforms

Snapchat, Instagram, ShareChat, and TikTok, among the leading players in the area of social media, have been notably proactive in adopting various solutions to maintain a conducive environment for their users. These solutions range from leveraging advanced AI for content moderation to implementing manual oversight and community-driven initiatives. In the following sections, we delve into the specific measures adopted by each platform to enhance user experience and safety.

1. **Snapchat**

   (a) **Family Center**: This feature is primarily a parental control tool tailored for teenagers aged 13 to 18. Parents can see with whom their teenagers have had conversations on Snapchat over the past week. However, the content of the messages remains inaccessible. The friend list of the teenager can also be accessed by parents to see who their child is connecting with. There is an avenue for parents to report concerns directly to Snapchat's Trust Safety team. Snapchat emphasizes that, by default, mutual friendship is necessary for communication and friend lists are kept private for safety. Parents can also filter out stories from publishers or creators marked as sensitive [59].
   (b) **Third-Party Apps (like SecureTeen)**: SecureTeen [60] provides a feature to block access to certain websites, apps, or content. It helps balance online and offline activities by allowing parents to set screen time limits. It offers insights into the child's social media interactions, including connections and shared content. It enables GPS tracking to monitor a child's location and offers features like

setting up geographical boundaries. It provides detailed reports on the child's online activities.

   (c) **Protection against 'My AI':** Snapchat has incorporated safety measures for the My AI feature, ensuring user-generated responses are not harmful or misleading. This includes steering clear of violent, hateful, sexually explicit, or otherwise offensive content [61].

2. **Instagram**

   (a) **AI and Machine Learning:** Instagram uses these technologies to detect potential content that breaches its Community Guidelines. This content is then reviewed by human moderators. The system also helps in processing user reports and blocking accounts that frequently violate platform policies [62, 63].

   (b) **Live Moderator:** This feature allows hosts of live streams to delegate moderation tasks, such as reporting comments or turning off comments for specific users [62].

   (c) **Hide Comments and Message Requests:** Users can automatically hide messages and comments that contain offensive content, ensuring a cleaner user experience. Users have the flexibility to customize filters and even hide specific words or emojis [64].

   (d) **Content Moderation Tools:** Platforms like Hootsuite, Respondology, and BrandFort assist in maintaining a positive environment on Instagram. Instagram's "nudity protection" feature aims to block unsolicited nude photos in direct messages using machine learning [65, 66].

3. **ShareChat**

   (a) **ShareChat Sakhi:** This initiative was launched to promote safety for women on the platform. Features such as blocking profiles, reporting issues, and restricting downloads/screenshots are available. ShareChat also created a dedicated email address for women to report any issues they face on the platform [67].

   (b) **Real-time Risk Intelligence with SHIELD:** ShareChat has integrated with SHIELD to provide real-time analytics, enhancing the platform's security [68].

   (c) **AI-Driven Content Moderation:** ShareChat uses AI to detect abusive content in multilingual text and multimodal content analysis, ensuring a clean platform. The feed algorithm is optimized using AI to ensure users aren't exposed to harmful content [69].

4. **TikTok**

   (a) **Content Moderation Approach:** TikTok balances automated technology and human review to ensure that content aligns with their Community Guidelines. This includes reviewing flagged content, popular content, and handling appeals from users [70].

   (b) **Community Reporting:** Users can report a plethora of issues, ranging from content violations to impersonation accounts, ensuring community involvement in platform safety [71].

   (c) **Community Guidelines:** TikTok's guidelines aim to maintain a safe and inclusive environment by removing violative content, age-restricting content, regulating goods and commercial activities, and more [72].
   (d) **Algorithm-Personalized Experience:** TikTok's algorithm considers user activity signals, engagement metrics, and other elements to tailor the content served to users. This ensures a balanced feed without inappropriate content.

5.1.2 AI Role in Content Moderation for social media apps

In this section, we will discuss how AI has been used efficiently by these platforms for content moderation processes.

1. **Snapchat:**
   (a) Automated Detection: Snapchat employs AI to swiftly spot and filter out content that violates its standards, from hate speech to explicit imagery.
   (b) Image and Video Analysis: Through AI, Snapchat can meticulously assess images and videos, preventing the sharing of inappropriate visuals on the platform.
   (c) Age Checks: With the help of AI operating over signals collected from user data, Snapchat verifies user ages to ensure content appropriateness, aligning with platform age restrictions.
   (d) Flagging Location-Based Risks: Snapchat, with AI's assistance, monitors real-time location data to detect potential threats, making Snap Map a safer feature for users.
   (e) Combatting Fake Identities: AI is crucial in identifying and blocking fake profiles and bots, ensuring a genuine user experience on Snapchat.

2. **Instagram**

   (a) Automated Detection: Advanced AI models have been deployed to swiftly identify explicit or inappropriate content, ensuring a safer user experience.
   (b) Image and Video Analysis: Through AI, Instagram can examine images and videos for sensitive or explicit material, thereby maintaining the platform's standards.
   (c) Age Verification: AI assists in ensuring the age-appropriateness of content, aligning with platform age restrictions.
   (d) Combatting Fake Identities: AI aids in the identification of fake profiles and bots, safeguarding against potential platform misuse.
   (e) Minimized Exposure: AI reduces the need for human intervention, minimizing their exposure to harmful content.
   (f) Streamlined Reporting: Automated reporting mechanisms powered by AI prioritize and escalate severe content violations.

3. **Sharechat**

   (a) Abuse Detection in Text: ShareChat has harnessed AI to detect and filter out abusive content present in textual formats. This technological approach enables the rapid identification and removal of objectionable content from the platform [73].

(b) Multimodal Content Analysis: AI is also utilized to analyze user-generated content spanning across different data modes such as text, images, and videos to flag inappropriate materials [69].

(c) Improved Datasets for Moderation: ShareChat introduced datasets like 3MAS-SIV for enhanced content understanding and MACD to detect abusive content in multiple Indian languages. The platform also released ADIMA for abuse detection in audio chatrooms [74–76].

(d) Optimized Feed Algorithm: AI has been instrumental in fine-tuning ShareChat's feed algorithm, ensuring users aren't exposed to any objectionable content.

(e) Video Content Moderation: ShareChat has developed AI-backed mechanisms to detect and remove Integrity Violating Content in videos [77].

TikTok

(a) Flagging Inappropriate Videos: AI systems scrutinize videos for content violations, including dangerous challenges, explicit scenes, and more.

(b) Monitoring Textual Content: AI-driven algorithms continuously oversee comments to detect threats, hate speech, or cyberbullying.

(c) Reviewing Reported Content: AI aids in efficiently handling large-scale reports by automatically reviewing flagged content.

(d) Analyzing Content Contextually: Beyond surface-level violations, AI delves deep to identify hidden policy violations based on the contextual analysis of videos.

(e) Streamlining Appeals: TikTok's appeals process for content removal is enhanced with AI, ensuring timely and fair resolutions.

5.2 The Gig Economy: Streaming and Gig Work Platforms

The rise of on-demand platforms has profoundly reshaped the economy and how we access online services and engage with service providers for physical and digital goods. Companies like Uber, DoorDash, and Twitch exemplify this digital transformation. Twitch, with 140 million monthly visitors (up from 55 million in 2015) and 107,800 live broadcasts at any time, wields significant influence, especially among the youth. Nearly 75 percent of its users are under 35, with 41 percent aged 16-24 and 32 percent aged 25-34. The US houses a quarter of all users, emphasizing Twitch's global impact. This dominance underscores Twitch's pivotal role in shaping youth culture, particularly in the realm of gaming. Uber has upended the transportation industry by providing a more personalized and hassle-free alternative to taxis. DoorDash has capitalized on surging demand for food delivery, especially in light of global events. Though differing in their offerings, these platforms collectively epitomize the paradigm shift towards digitally-mediated, on-demand, and user-focused consumption. They cater to modern preferences for customizable, instant and targeted services and entertainment. Their meteoric success underscores how digital innovation can disrupt traditional business models to better serve evolving consumer needs and behaviors. We examine the problems faced by specific platforms, substantiating them with evidence, to offer a comprehensive overview of the pressing challenge.

1. **Twitch**

Category	App	Panic Button/SOS	Blocking	Reporting
Dating Apps	**Tinder**	allows users to discretely summon emergency services if threatened on dates	Instantly cease contact with inappropriate or abusive matches	address inappropriate behavior and content
	Bumble	Same As Above	Same As Above	Same As Above
	OkCupid	NA	Same As Above	Same As Above
Category	**App**	**Panic Button/SOS**	**Background Verification**	**Reporting**
Gig and Streaming Platforms	**Uber**	riders and drivers to secretly call for assistance when facing real-world danger,	Comprehensive driver background checks mitigate risks from high-contact roles	Streamlined reporting systems empower riders and drivers to discreetly escalate issues
	Twitch	NA	NA	flag inappropriate behavior for swift enforcement
	Doordash	delivery workers urgently alert emergency services if threatened	Rigorous screening of delivery workers through background checks enhances safety for customers	Easy safety reporting functionality lets delivery workers and customers raise concerns
Category	**App**	**Moderators**	**Blocking**	**Reporting**
Social Media	**Snap**	ensure nuanced policy enforcement not fully addressed by AI	stop unwanted interactions	flagging and resolving violations
	Instagram	Same as Above	Blocking abusive accounts/comments helps curb toxicity	Same as Above
	Sharechat	Same as Above	Same as Above	Same as Above
	TikTok	Same as Above	helps restrict cyberbullying and harassment	flagging and resolving violations

Table 4 Platform-specific Safety Solutions and Measures: This table delineates the various safety mechanisms implemented across different digital platforms, from dating apps to social media. Whether through background checks for gig workers or through streamlined reporting systems, these platforms are taking crucial steps to ensure a more secure user experience.

(a) Issues with Abusive Behavior: Streamers and viewers often engage in racist, sexist, homophobic, or abusive behaviors. Twitch's own transparency reports indicate numerous violations related to hate speech and harassment. High-profile streamers like Ice Poseidon and Greekgodx have been involved in numerous incidents highlighting racism, sexism, and homophobia. Marginalized streamers have also come forward discussing the harassment they face, and the platform has faced advertiser boycotts linked to its toxicity issues [78, 79].

(b) Dissemination of Mature/Explicit Content: The user-driven content creation nature of Twitch leads to the dissemination of objectionable or explicit materials. Although Twitch's policies prohibit such content, the platform struggles to control nudity, pornography, and extreme violence in streams. The gaming content on Twitch, which frequently contains elements like violence, gore, and drug/alcohol depictions, is often a challenge when considering what is acceptable.

(c) Malicious Links in Chats/Streams: The clicking of malicious links shared in chat or during a stream. Research from Symantec has highlighted the use of Twitch chats for distributing spam and malware. Scams involving fake free gift card links have become prevalent, causing financial and data-related harm to users [80].

2. **Uber**

(a) Fake or Duplicate User Accounts: The use of fake or duplicate accounts for purposes like fraud and spamming. here have been identified patterns like inconsistent or suspicious ratings, multiple account registrations from a single device or IP, and the usage of identical email addresses or phone numbers for various accounts.

(b) Scams, Phishing, and Malicious Links: Users might be subjected to scams, phishing attempts, or other malicious links shared through user messages or profiles. Users have flagged certain profiles for suspicious behavior, like asking for personal or financial details. There have also been incidents where riders or drivers received questionable messages or links from other users [81].

(c) Harassment, Stalking, and Unwanted Messaging: Harassment, stalking, and other unwanted forms of communication between drivers and riders. Instances have been recorded, ranging from chat conversations that contain inappropriate messages to voice recordings that capture episodes of harassment [82].

3. **DoorDash**

(a) Phishing and Fraudulent Communications: DoorDash users might encounter unsolicited communications that falsely claim to be from the platform. These messages, often via text, email, or phone calls, might be phishing attempts, trying to deceive users into divulging personal details or engaging with harmful links.

(b) Phishing Sites and Data Theft: Users are at risk when clicking on links from dubious sources, as these might redirect them to phishing sites that are structured to pilfer login data, personal details, or payment information.

(c) Account Compromise: Due to weak or reused passwords, users' DoorDash accounts may be vulnerable to hacking. Once these unauthorized individuals gain access, they can misuse the stored personal information.

5.2.1 Safety measures taken by these platforms

Companies like Twitch, Uber, and DoorDash have become integral parts of our daily lives, making it even more crucial for them to prioritize the safety and security of their user base. In this section, we delve into the safety measures these platforms have integrated to create a safer ecosystem for their users.

1. **Twitch**

(a) Addressing Abusive Behaviour: Twitch allows both streamers and users to report abusive chats, profiles, and streams either via on-site reporting mechanisms or through a dedicated email. For severe breaches of its community guidelines, Twitch enforces stringent policies like permanent bans. Repeated infractions can result in account suspensions. Twitch deploys a mix of user tools, moderator functions, strict policy enforcement, and product features to tackle harassment. To defend against targeted attacks, users have the option to activate the 'shield mode', which lets them instantly customize and activate safety protocol [83].

(b) Tackling Mature/Explicit Content: Twitch has set clear regulations around sexual content, age-restricted games, and prohibited games. Streamers, including some popular ones, have been banned due to violations such as sexual exploitation and hate conduct [84].

(c) Combating Malicious Links: Twitch proactively screens for prohibited URL domains that are known affiliates of scams and spam. Some chat rooms restrict posting links only to moderators or verified accounts. Twitch moderators rigorously review reports of suspicious links and take corrective actions like removing the links or suspending the offending accounts. Accounts that solely spam links are also purged to mitigate malicious traffic [85].

2. **Uber**

(a) Mitigating Fake or Duplicate Accounts: Uber mandates new users to furnish personal details including phone numbers, emails, and payment information, which are verified prior to granting access. The verification process for drivers is even more rigorous. Features like sending a one-time password to a user's phone during login adds an additional security layer. Uber employs dedicated anti-fraud teams that manually inspect suspicious accounts using specialized tools [86].

(b) Addressing Scams and Phishing: Uber maintains a list of blocked keywords and domains to auto-filter known scam or phishing links. Specialized teams are in place to investigate user complaints related to scams or harassment. Uber deploys software to identify accounts that might be partaking in fraudulent activities. Every suspicious activity is manually reviewed by a specialized team before confirming its fraudulent nature. [87]

(c) Tackling Harassment: Uber allows mutual blocking of users, ensuring they never get matched again. The review system permits riders and drivers to anonymously share feedback about unsavory experiences. Those found guilty of harassment may face account deactivation. All driver applicants undergo thorough background checks [88].

3. **DoorDash**

(a) SafeChat: Designed to foster a respectful and safe environment, SafeChat springs into action when inappropriate or offensive language is detected in a chat. The sender is issued a warning while the receiver–if a Dasher–can opt out of the delivery without any repercussions and also has the option to report the issue [89].

(b) Safety Reporting: Dashers can report unsettling interactions with customers immediately via in-app chat or call. They can also block future deliveries from the concerned customer [90].

(c) Real-Time Safety Alerts: Developed in collaboration with samdesk [6], this feature apprises dashers, customers, and merchants of real-time emergencies or disasters. It even cancels ongoing deliveries in affected areas [91].

(d) OPA (Open Policy Agent): To boost developer efficiency, DoorDash integrated Open Policy Agent. The infrastructure team reported several benefits like faster

[6] https://www.samdesk.io/

reviews of infrastructure policy changes, comprehensive resource tagging, and a significant reduction in incidents resulting from policy breaches [92].

5.2.2 Content Moderation with AI

By virtue of interpersonal interaction in gig work applications as well as user-generated real-time content on streaming platforms, there are unique challenges around content moderation. AI can be deployed in various internal and external-facing applications to support them in addressing such challenges. Here are some of the opportunities for such a deployment, many of which are actively being explored.

1. **Twitch**
   - (a) Abusive Behaviour on Twitch: Natural Language Processing (NLP) can be harnessed to identify abusive speech, hate speech, threats, and more within chat and user profiles. Sentiment analysis has the potential to pinpoint hostility, aggression, or bullying in chat messages. By training image and video classifiers, the platform can recognize objectionable gestures, scenes, and behavior in live streams. Object detection models can be tasked with pinpointing imagery, attire, or activities that breach Twitch's policies. Recommender systems can suggest measures like auto-bans, blocks, and timeouts for accounts that consistently flout community guidelines.
   - (b) Dissemination of Mature/Explicit Content: NLP algorithms can be used to flag profanity, sexual words, and drug references in chat, user profiles, and stream metadata. Game footage can be categorized as either permissible or age-restricted through visual content analysis. To ensure streams are within Twitch's guidelines, AI can detect nudity, violence, and gore. Streamed games can be compared against a platform database to assess compliance with age-rating policies. Thumbnails, overlays, and streamer attire can be moderated using sophisticated image classifiers.
   - (c) Malicious Links: URL reputation databases and real-time web scraping algorithms can identify links associated with scams or spam. Classifiers can be trained to recognize patterns indicative of bot accounts spamming identical links. Historical data analysis can highlight accounts likely to be spam bots. Shared links can be vetted against lists of blocked domains. Phishing attempts and other forms of social engineering can be identified through NLP.
2. **Uber**
   - (a) Fake or Duplicate User Accounts: AI-assisted verification can be applied to identity documents and vehicle papers during the registration process. Account connections and user clusters can be analyzed to detect groups of counterfeit accounts. Behavioral patterns in account actions (such as rides and ratings) can be monitored to pinpoint accounts that diverge from typical user activity. Language models can help differentiate genuine users from bots or scammers based on communication styles.
   - (b) Scams, Phishing, or Other Malicious Links Shared: AI can be employed to rapidly and efficiently scan textual, visual, and audio content for potentially

harmful elements. Communication patterns can be analyzed to detect spam campaigns or coordinated scams. Differences in language and communication styles can help discern genuine users from potential fraudsters.

   (c) Harassment, Stalking, and Unwanted Messaging between Drivers and Riders: Offensive language, aggression, and inappropriate advances in messages can be flagged by AI models. Hypothetical conversational scenarios can be generated to aid in the training of both AI models and human moderators, ensuring they can effectively identify harassment or stalking. Audio messages can be scrutinized for inappropriate or threatening content using specialized algorithms.

3. **DoorDash**

   (a) SafeChat - Automated Message Monitoring: A standout feature in DoorDash's AI arsenal is SafeChat. This automated system monitors messages exchanged within the app, specifically focusing on detecting any offensive or abusive language. If such content is identified, the system immediately issues a warning to the sender, irrespective of whether they're a customer or a DoorDash associate [93].

   (b) SafeDash Check-in - Proactive Safety Alerts: Expanding on its previously introduced safety toolkit, SafeDash, DoorDash has launched the SafeDash Check-in. This feature is inherently proactive, designed to automatically check in with delivery workers if the system detects unusually prolonged delivery times. If a response isn't received from the worker within a stipulated time, an ADT agent reaches out to provide assistance, ensuring their safety [93].

   (c) Porch Light Reminders: Understanding the challenges faced by delivery personnel during nighttime deliveries, DoorDash prompts customers to turn on their outdoor lights, simplifying address identification and augmenting safety [93].

5.3 Dating Apps

With technology-driven relationships, dating apps have become indispensable tools for driving connections among youth audiences. With the surge in online dating, platforms like Bumble, Tinder, and OkCupid play pivotal roles in the romantic journeys of young adults. Each platform offers unique features tailored to cater to its user base. Bumble, with a median age of 26, has garnered significant popularity among the younger generation [94]. The app's design, which allows women to initiate conversations, offers an element of control and empowerment. By providing age range filters, users can streamline potential matches, ensuring that their dating pool aligns with their age preferences. Bumble's commitment to chivalry and its fast-paced nature, given its 24-hour response window, make it particularly enticing for the youth. The platform's former feature that facilitated a dedicated space for teenagers further exemplified Bumble's understanding of the youth's dating needs. Tinder, predominantly favored by younger audiences, is used by 79% of online daters under 30 [13]. The platform stands out for its instantaneous gratification, allowing users to quickly swipe, match, and engage. Primarily used for seeking romantic connections, Tinder also serves as an entertainment platform. Its swiping mechanism offers an enjoyable user experience,

Category	App	Toxic/offenisive text detection	NSFW multimedia Detector	AI-based profile Verification
Dating Apps	**Tinder**	curbing harassment and abuse in messages	blocking unsolicited explicit images and other multimodal media	Selfie based photo verification to maintain authenticity on platform
	Bumble	Same As Above	Same As Above	Same As Above
	OkCupid	Sams As Above	Same As Above	NA
Category	**App**	**Toxic/Offensive text detection**	**Fake Profile Verification**	**Spam Detection**
Gig and Streaming Platforms	**Uber**	curb toxicity between rider and customer	To protect riders or customers from harassment and frauds	To stop promotional spam and phishing attempts
	Twitch	curb toxicity in chatrooms	to protect the identity of streamers	Same as Above
	Doordash	curb toxicity between customer and delivery merchant	To protect delivery merchants or customers from harassment and frauds	Same as Above
Category	**App**	**Toxic/Offensive text detection**	**NSFW Detector**	**Bot Detection**
Social Media	**Snap**	Curb abuse and toxicity in messages	blocking unsolicited explicit images and other multimodal media	To remove fake accounts
	Instagram	Same as Above	Same as Above	To remove accounts that run influence operations and manipulate ranking feeds
	Sharechat	Same as Above	Same as Above	Same as Above
	TikTok	Same as Above	helps restrict cyberbullying and harassment	flagging and resolving violations

Table 5 AI-powered Solutions Across Digital Platforms: This table highlights the integration of AI-driven features to combat prevalent challenges across various online platforms, from dating apps to social media.

and matching provides a confidence boost, giving users a sense of social validation. Its perceived maturity draws younger users seeking to emulate adult behavior. OkCupid, while not as predominantly used by Gen Z as the other platforms, offers unique features that appeal to a broader audience. The platform boasts an expansive array of questions that feed its matching algorithm, ensuring that matches are based on shared values and interests. Significantly, OkCupid has taken leaps in inclusivity by introducing 22 genders and 13 orientations, allowing users to genuinely express their identities. This progressive stance has made OkCupid an attractive platform for those valuing authentic self-representation. As digital courtship evolves, dating apps are continuously refining their offerings to cater to the dynamic needs of today's youth. Whether through empowerment, instant gratification, or inclusivity, these platforms provide varied avenues for young individuals to navigate the complex world of romance. We examine the problems faced by specific platforms, substantiating them with evidence, to offer a comprehensive overview of the pressing challenge:

1. **Bumble**

   (a) Cyber-flashing: Unsolicited explicit images are often sent to users, leading to emotional distress and potential violation of privacy [95].

(b) Sexual Manipulation: Young adults can be at risk of grooming, exploitation, and manipulation by predatory users [96].

(c) Underage Use: The potential exists for underage users to access Bumble, exposing them to risk [97].

(d) General Risks: Young adults face the usual dangers seen on other dating platforms, including exposure to violent predators [98].

(e) Privacy Concerns: Bumble's collection of personal data can be at risk of unauthorized access and misuse.

2. **Tinder**

(a) Predators: Teenagers may encounter individuals with malicious intent, aiming to groom or exploit them.

(b) Extortion Risks: Sharing intimate photos can lead to threats of public exposure.

(c) Privacy Concerns: The vast amount of data Tinder collects could be potentially accessed by hackers or cybercriminals.

(d) Stalking and Harassment: Users have reported stalking and harassment from matches, leading to psychological distress.

(e) Cyberbullying: The platform can be a breeding ground for bullying, especially for younger users.

3. **OkCupid**

(a) Security Flaws: Past vulnerabilities could have let hackers access users' personal data like messages and sexual orientation [99, 100].

(b) No Age Verification: Lack of a robust age verification system allows potentially underage users to access the platform.

5.3.1 Safety Measures

Popular dating apps like Bumble, Tinder, and OkCupid have incorporated various safety measures to protect their community and enhance user experience. This subsection delves into the specific features these platforms have implemented to ensure a secure environment for their users.

1. **Bumble**

(a) Weapons Ban: Bumble ensures a more safe environment by prohibiting guns and other violent weapons in profile pictures.

(b) Communication Security: Users can video chat and voice call directly within the Bumble app, avoiding the need to share personal contact information prematurely.

(c) Mental Well-being: The Snooze feature lets users temporarily pause their profiles if they feel the need to take a break from dating.

(d) Blocking Reporting: With the Unmatch and Block Report systems, users can easily avoid or report violators of Bumble's community guidelines.

(e) Safety Guide: Bumble has launched a ”Stand for Safety” initiative with a women's safety guide to combat online abuse [101].

(f) Hate Speech and Misbehavior: The platform promotes a positive environment, discouraging hate speech, bullying, and misogyny.

(g) Photo Verification: Users' identities are confirmed through photo verification to prevent catfishing.

(h) Private Detector: An automatic blur on inappropriate images, giving users control over viewing explicit content.

(i) Safety IRL: Provides users with practical safety guidelines for real-life dates.

2. **Tinder**

   (a) Message Screening: Advanced machine learning tools detect and warn users of potentially offensive messages.

   (b) Guidance Features: Features like Just-In-Time guidance and alerts for LGBTQ+ users in unfriendly regions promote user awareness and protection.

   (c) Long Press Reporting: A more seamless way to report inappropriate messages by pressing and holding the offensive text.

   (d) Panic Button: In case of real-life threats during dates, users can alert emergency services using the panic button.

   (e) Photo Verification: This feature helps in authenticating the identity of users to reduce fake profiles.

   (f) Background Checks: An upcoming feature allowing users to conduct background checks on their matches.

3. **OkCupid**

   (a) Reporting: Users can easily report suspicious or inappropriate activities to maintain community standards.

   (b) Safety Guidelines: OkCupid provides essential safety advice, emphasizing personal and financial security during interactions [102].

5.3.2 Content Moderation with AI

The integration of AI into dating apps has revolutionized content moderation, making it more efficient, real-time, and adaptive. With millions of users interacting daily, manual moderation becomes an arduous task. AI steps in to automate the detection and prevention of inappropriate content, abusive behavior, and other violations of platform guidelines. Let's examine how prominent dating apps like Bumble, OkCupid, and Tinder leverage AI to ensure a safer user experience.

1. **Bumble**

   (a) 'Private Detector' feature: Bumble's AI-driven tool helps shield users from unsolicited photos, including explicit content, shirtless selfies, and firearm images. Images deemed inappropriate are automatically blurred, allowing the recipient the discretion to view, block, or report. The underlying algorithm boasts a 98% accuracy in detecting such content, utilizing an EfficientNetv2-based binary classifier.

(b) Image Verification Tool: Bumble combats fake profiles through its AI-powered image verification system. By mandating users to mimic random poses, Bumble's AI compares these selfies with profile images to authenticate users.

(c) AI Tools for Flagging: AI assists Bumble's moderation process by autonomously highlighting suspicious or inappropriate photos. Such flagged profiles remain concealed until user verification.

(d) Multilingual Language Models: Bumble employs large language models capable of detecting unwanted messages in multiple languages. These models understand intricate nuances across languages, ensuring a respectful interaction environment [103].

2. **Tinder**

(a) Monitoring Conversations: Tinder's AI continually assesses direct messages, flagging inappropriate or harmful content.

(b) "Are You Sure?" Feature: A proactive tool, this AI-driven feature scans private conversations for potentially offensive language, prompting users to reconsider their message.

(c) Profile Verification: Tinder utilizes AI to verify the authenticity of user profiles, reducing the prevalence of bots or catfish accounts. This verification juxtaposes user photos with official ID images, flagging discrepancies for further review.

(d) Facial Identification: AI-driven facial recognition confirms user authenticity by comparing live selfies with profile images.

3. **OkCupid**

(a) Machine Learning Alerts: OkCupid's support system uses machine learning to instantaneously detect harmful or abusive language, empowering the moderation team to act swiftly.

6 Limitations of AI based Content Moderation Techniques

The application of AI in the realm of content moderation across diverse platforms, from social media to dating apps, shows great promise but is not without its pitfalls. In the sphere of social media, we have decades-old examples of technological interventions being abused as notoriously exemplified by Microsoft's chatbot Tay, the risks of AI's uncontrolled learning from unfiltered data were laid bare. Tay began to make offensive remarks after being influenced by toxic user interactions, revealing a pressing need for greater safeguards when it comes to AI operating freely online [104]. Similarly, Facebook's AI content filtering system missed moderating the New Zealand mosque shooting video, pointing to a shortfall in its training data which lacked sufficient violent attack examples [105]. The inherent dangers of this oversight emphasize the need to delve deeper than just surface-level content and genuinely understand the nuances of videos. Twitter also fell foul when its photo cropping algorithm demonstrated racial bias, shedding light on the limitations of non-diverse data and the lack of model transparency [106].

Streaming and gig platforms, such as Twitch, haven't been immune to the challenges of AI moderation either. A recurrent issue is AI-based models struggling to decipher visuals correctly, given its inability to appreciate the broader context in which they are set. This becomes a significant problem when malicious users craft content to bypass established patterns, evading AI detection. Platforms like Uber face a different challenge, where the AI, if trained on biased data, could perpetuate societal prejudices, leading to discriminatory practices.

On the one hand, platforms like Tinder integrate AI to refine user experiences, making them safer and more genuine. However, one of the primary concerns users have is related to privacy. The idea that AI is scanning and analyzing private conversations, despite being in the pursuit of identifying inappropriate content, has given rise to significant discomfort. This monitoring, even if well-intended, often borders on intrusive. Further, with the algorithms behind these AI tools being kept under wraps by major dating platforms, users are left with a plethora of unanswered questions about how their data is used and analyzed. The adaptability of AI, while impressive, is a double-edged sword. It can be harnessed for misuse, with features like auto-swiping and messaging potentially leading to an increase in fake profiles and spam. Moreover, the potential for AI to misconstrue human nuances, like sarcasm or cultural banter, can result in unwarranted content flags. Additionally, if AI models are trained on data sets that carry inherent biases, it can lead to unjust content flagging based on aspects like race or gender.

In a nutshell, while AI offers a transformative edge to content moderation across platforms, its current limitations underline the importance of human oversight, diverse training data, and clear methodologies in its ethical and effective use.

7 Conclusion

There are many common challenges that youth-social platforms deal with in their day to day operations; most of which are likely to be exacerbated by the advent of GenAI. Broadly, there are content-related and actor-related harms that are widely prevalent and early efforts indicate the need for a hybrid AI driven filtering combined with human adjudication prior to enforcement. It is incredibly challenging to deal with these harms in siloes, which is at present how most platforms tend to deal with harms, due to a lack of appropriate data and policy sharing mechanisms. With regulatory requirements requiring platforms to pursue additional contingencies including appeals processes for users, transparency requirements, research access to datasets, we find that it would be beneficial to also define governance structures that permit the sharing of successes and failures with regards to content moderation, policy decisions, and platform design choices. AI driven content moderation is not likely to be the panacea given its limitations and the inexcusable cost of false positives in end-to-end automation for enforcement decisions. It is also true that most vulnerable populations tend to have the least safeguards designed due to the sociocultural complexities of designing bespoke safety measures and policies for them. This undergirds our proposition for regulation to support additional policy impact sharing mechanisms and data, similar to the 'blameless post-mortems' followed in the field of software engineering that has resulted in industry-wide best practices to transparently share incidents and increase platform awareness of harms, ultimately benefitting the end users. It is easier to make progress on challenging problems afflicting platforms through collaborative efforts, as has been repeatedly shown with the creation of popular benchmarks like the Netflix Recommendation challenge, the ImageNet challenge, and recently, open-source large language models and diffusion models. Additionally, sourcing vulnerabilities and harms allows more public visibility and awareness of the pitfalls of nascent technologies. This review provides evidence to justify that there is a clear need for such sharing structures, with viable benefits for involved platforms, regulators, and most importantly, the youth using these platforms actively as part of their digital experiences.

8 Conflict of Interest Statement

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

[1] Luca, M.: User-generated content and social media. In: Handbook of Media Economics vol. 1, pp. 563–592. Elsevier, ??? (2015)

[2] Name, A.: Types of AI content moderation and how they work. TechTarget. Accessed: dd-mm-yyyy (2023). https://www.techtarget.com/searchcontentmanagement/tip/Types-of-AI-content-moderation-and-how-they-work

[3] Name, A.: Automated Content Moderation: A Primer. Stanford Cyber Policy Center. Accessed: dd-mm-yyyy (2023). https://cyber.fsi.stanford.edu/news/automated-content-moderation-primer

[4] Name, A.: Using GPT-4 for Content Moderation. OpenAI. Accessed: dd-mm-yyyy (2023). https://openai.com/blog/using-gpt-4-for-content-moderation

[5] Zhou, X., Zhu, H., Yerukola, A., Davidson, T., Hwang, J.D., Swayamdipta, S., Sap, M.: COBRA frames: Contextual reasoning about effects and harms of offensive statements. In: Findings of the Association for Computational Linguistics: ACL 2023, pp. 6294–6315. Association for Computational Linguistics, Toronto, Canada (2023). https://doi.org/10.18653/v1/2023.findings-acl.392 . https://aclanthology.org/2023.findings-acl.392

[6] Gillespie, T.: Content moderation, ai, and the question of scale. Big Data & Society **7**(2), 2053951720943234 (2020) https://doi.org/10.1177/2053951720943234 https://doi.org/10.1177/2053951720943234

[7] Name, A.: Catching Bad Content in the Age of AI. MIT Technology Review. Accessed: dd-mm-yyyy (2023). https://www.technologyreview.com/2023/05/15/1073019/catching-bad-content-in-the-age-of-ai/

[8] Name, A.: AI Large Language Models, Data Scraping, and Generation: Remaking the Web. The Verge. Accessed: dd-mm-yyyy (2023). https://www.theverge.com/2023/6/26/23773914/ai-large-language-models-data-scraping-generation-remaking-web

[9] Researchers Discover a Way to Make ChatGPT Consistently Toxic. TechCrunch (2023). https://techcrunch.com/2023/04/12/researchers-discover-a-way-to-make-chatgpt-consistently-toxic/

[10] Pan, Y., Pan, L., Chen, W., Nakov, P., Kan, M.-Y., Wang, W.Y.: On the risk of misinformation pollution with large language models. arXiv preprint arXiv:2305.13661 (2023)

[11] 7 Facts About Americans and Instagram. Pew Research Center (2021). https://www.pewresearch.org/short-reads/2021/10/07/7-facts-about-americans-and-instagram/

[12] Teens, Social Media, and Technology 2022. Pew Research Center (2022). https://www.pewresearch.org/internet/2022/08/10/teens-social-media-and-technology-2022/

[13] Key Findings About Online Dating in the U.S. Pew Research Center (2023). https://www.pewresearch.org/short-reads/2023/02/02/key-findings-about-online-dating-in-the-u-s/

[14] Americans' Experiences Earning Money Through Online Gig Platforms. Pew Research Center (2021). https://www.pewresearch.org/internet/2021/12/08/americans-experiences-earning-money-through-online-gig-platforms/

[15] Social Media Use Continues to Rise in Developing Countries but Plateaus Across Developed Ones. Pew Research Center (2018). https://www.pewresearch.org/global/2018/06/19/social-media-use-continues-to-rise-in-developing-countries-but-plateaus-across-developed-ones/

[16] Bickert, M.: Charting a Way Forward: Online Content Regulation. White Paper, Facebook (2020). available at : https://about.fb.com/wp-content/uploads/2020/02/Charting-A-Way-Forward_Online-Content-Regulation-White-Paper-1.pdf

[17] Snapchat: Transparency Report. Privacy and Safety Hub. available at : https://values.snap.com/privacy/transparency

[18] Snapchat: Government Requests and Copyrighted Content Takedown Notices (DMCA). Privacy and Safety Hub. available at : https://values.snap.com/privacy/transparency/legal-requests

[19] Google YouTube: Information quality and content moderation. Technical report, Google (2021). available at : https://blog.google/documents/83/information_quality_content_moderation_white_paper.pdf/

[20] Google YouTube: A new policy on advertising for speculative and experimental medical treatments. Technical report, Google (2019). available at : https://support.google.com/google-ads/answer/9475042?hl=en

[21] Google YouTube: Threat analysis group. Technical report, Google (2018). available at : https://blog.google/threat-analysis-group/

[22] Mishra, A.: Analysis of Social Media Compliance Reports for the Month of April 2022. Internet Freedom Foundation. available at : https://internetfreedom.in/tag/socialmediacompliancewatch/

[23] Pardes, A.: Tinder Asks 'Does This Bother You'? Wired. available at : https://www.wired.com/story/tinder-does-this-bother-you-harassment-tools/

[24] Tik Tok: Strengthening Our Policies to Promote Safety, Security, and Well-being on TikTok. Safety. available at : https://www.tiktok.com/transparency/en-us/community-guidelines-enforcement/?enter_method=category_card

[25] New AI Classifier for Indicating AI-Written Text. OpenAI (2023). https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text

[26] Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., Goldstein, T.: A Watermark for Large Language Models (2023)

[27] Kirchenbauer, J., Geiping, J., Wen, Y., Shu, M., Saifullah, K., Kong, K., Fernando, K., Saha, A., Goldblum, M., Goldstein, T.: On the Reliability of Watermarks for Large Language Models (2023)

[28] Watanabe, H., Bouazizi, M., Ohtsuki, T.: Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. IEEE access **6**, 13825–13835 (2018)

[29] Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 11, pp. 512–515 (2017)

[30] Paul, S., Saha, S.: Cyberbert: Bert for cyberbullying identification. Multimedia Systems, 1–8 (2020)

[31] Kamble, S., Joshi, A.: Hate speech detection from code-mixed hindi-english tweets using deep learning models. arXiv preprint arXiv:1811.05145 (2018)

[32] Tabone, A., Camilleri, K., Bonnici, A., Cristina, S., Farrugia, R., Borg, M.: Pornographic content classification using deep-learning. In: Proceedings of the

21st ACM Symposium on Document Engineering. DocEng '21. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3469096.3469867 . https://doi.org/10.1145/3469096.3469867

[33] George, E., Surdeanu, M.: It is not Sexually Suggestive, It is Educative. Separating Sex Education from Suggestive Content on TikTok Videos (2023)

[34] Micallef, N., Sandoval-Castañeda, M., Cohen, A., Ahamad, M., Kumar, S., Memon, N.: Cross-platform multimodal misinformation: Taxonomy, characteristics and detection for textual posts and videos. Proceedings of the International AAAI Conference on Web and Social Media **16**(1), 651–662 (2022) https://doi.org/10.1609/icwsm.v16i1.19323

[35] Yao, B.M., Shah, A., Sun, L., Cho, J.-H., Huang, L.: End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '23, pp. 2733–2743. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3539618.3591879 . https://doi.org/10.1145/3539618.3591879

[36] He, B., Ahamad, M., Kumar, S.: Reinforcement Learning-based Counter-Misinformation Response Generation: A Case Study of COVID-19 Vaccine Misinformation (2023)

[37] Qian, J., Bethke, A., Liu, Y., Belding, E., Wang, W.Y.: A benchmark dataset for learning to intervene in online hate speech. arXiv preprint arXiv:1909.04251 (2019)

[38] Saha, P., Singh, K., Kumar, A., Mathew, B., Mukherjee, A.: Countergedi: A controllable approach to generate polite, detoxified and emotional counterspeech. arXiv preprint arXiv:2205.04304 (2022)

[39] Tran, M., Zhang, Y., Soleymani, M.: Towards a friendly online community: An unsupervised style transfer framework for profanity redaction. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 2107–2114. International Committee on Computational Linguistics, Barcelona, Spain (Online) (2020). https://doi.org/10.18653/v1/2020.coling-main.190 . https://aclanthology.org/2020.coling-main.190

[40] Meta Launches New Content Moderation Tool. Meta (2022). https://about.fb.com/news/2022/12/meta-launches-new-content-moderation-tool/

[41] Help Article. Facebook (2023). https://www.facebook.com/help/1584908458516247

[42] The Role of Artificial Intelligence in Snap: Transforming the Future of Social Media. Medium (2023). https://medium.com/@sukhveersinghdhiman74/

the-role-of-artificial-intelligence-in-snap-transforming-the-future-of-social-media-1b105f3b06fb

[43] Content Moderation and the Language of Artificial Intelligence (2023)

[44] Essential TikTok Stats. DataReportal (Year Not Provided). https://datareportal.com/essential-tiktok-stats

[45] Essential Instagram Stats. DataReportal (Year Not Provided). https://datareportal.com/essential-instagram-stats

[46] Snapchat Statistics. The Social Shepherd (Year Not Provided). https://thesocialshepherd.com/blog/snapchat-statistics

[47] Bayer, J., Ellison, N., Schoenebeck, S., Falk, E.: Sharing the small moments: Ephemeral social interaction on snapchat. Information Communication and Society **19**, 956–977 (2016) https://doi.org/10.1080/1369118X.2015.1084349

[48] Snapchat's new map feature raises fears of stalking and bullying. The Guardian (2017). https://www.theguardian.com/technology/2017/jun/23/snapchat-maps-privacy-safety-concerns

[49] Safety Concerns for Snapchat's New Snap Map Feature. Cybersmile Foundation (2023). https://www.cybersmile.org/news/safety-concerns-for-snapchats-new-snap-map-feature

[50] Instagram Statistics. Hootsuite (2023). https://blog.hootsuite.com/instagram-statistics/

[51] The Underworld of Online Content Moderation. The New Yorker (2023). https://www.newyorker.com/news/q-and-a/the-underworld-of-online-content-moderation

[52] Fighting deepfakes: When detection fails. Brookings Institution (2023). https://www.brookings.edu/articles/fighting-deepfakes-when-detection-fails/

[53] How regional social media platforms spew fake news and get away with it. Hindustan Times (2023). https://www.hindustantimes.com/opinion/how-regional-social-media-platforms-spew-fake-news-and-get-away-with-it/story-s8Kc2s4TKfne0ZRlXNuLuM.html

[54] ShareChat leans on AI to expand regional language user base. TechRadar (2023). https://www.techradar.com/news/sharechat-leans-on-ai-to-expand-regional-language-user-base

[55] Dangerous Social Media Challenges. FamilyMinded (2023). https://www.familyminded.com/s/dangerous-social-media-challenges-31c1b2b2be14413b

[56] Why Do People Participate in Dangerous Viral Challenges? Verywell Mind (2023). https://www.verywellmind.com/why-do-people-participate-in-dangerous-viral-challenges-5200238

[57] Music Labels Increase Detection of Unlicensed Music Use in Social Posts. Social Media Today (2023). https://www.socialmediatoday.com/news/Music-Labels-Increase-Detection-of-Unlicensed-Music-Use-in-Social-Posts/638382/

[58] How TikTok's algorithm exploits the vulnerability of children. The Guardian (2023). https://www.theguardian.com/technology/2023/apr/04/how-tiktoks-algorithm-exploits-the-vulnerability-of-children

[59] Introducing Content Controls on Family Center. Snap Inc. (2023). https://values.snap.com/news/introducing-content-controls-on-family-center

[60] Snapchat Parental Control. SecureTeen (2023). https://secureteen.com/parental-control/snapchat-parental-control

[61] What is My AI on Snapchat and how do I use it? Snapchat Help Center (2023). https://help.snapchat.com/hc/en-us/articles/13266788358932-What-is-My-AI-on-Snapchat-and-how-do-I-use-it-

[62] Help Article. Instagram Help Center (2023). https://help.instagram.com/477434105621119

[63] AI Content Moderation. Label Your Data (2023). https://labelyourdata.com/articles/ai-content-moderation

[64] Help Article. Instagram Help Center (2023). https://help.instagram.com/700284123459336

[65] Instagram's Stricter Content Moderation on Nudity. Yahoo News (2023). https://www.yahoo.com/now/instagram-stricter-content-moderation-nudity-115011081.html

[66] Content Moderation. Hootsuite (2023). https://blog.hootsuite.com/content-moderation/

[67] ShareChat introduces ShareChat Sakhi, a Women's Safety Feature. Social Samosa (2020). https://www.socialsamosa.com/2020/03/sharechat-sharechat-sakhi-womens-safety-feature/

[68] Case Study: ShareChat. Shield (2023). https://shield.com/case-studies/sharechat

[69] Multimodal Automated Content Moderation. ShareChat on Medium (2023). https://medium.com/sharechat-techbyte/

multimodal-automated-content-moderation-69876e6a9d85

[70] Content Moderation Transparency. TikTok (2023). https://www.tiktok.com/transparency/en-us/content-moderation/

[71] Evolving our approach to content enforcement. TikTok Newsroom (2023). https://newsroom.tiktok.com/en-eu/evolving-our-approach-to-content-enforcement

[72] TikTok Community Guidelines. TikTok (2023). https://www.tiktok.com/community-guidelines/en/

[73] Dhinakaran, A.: Can AI Help Make Social Media More Accessible, Inclusive, And Safe? Forbes (2021). https://www.forbes.com/sites/aparnadhinakaran/2021/12/14/can-ai-help-make-social-media-more-accessible-inclusive-and-safe/?sh=5b308edcf6cb

[74] Adima: A Benchmark for Automated Moderation. ShareChat (2023). https://sharechat.com/research/adima

[75] Gupta, V., Mittal, T., Mathur, P., Mishra, V., Maheshwari, M., Bera, A., Mukherjee, D., Manocha, D.: 3MASSIV: Multilingual, Multimodal and Multi-Aspect dataset of Social Media Short Videos (2022)

[76] Gupta, V., Roychowdhury, S., Das, M., Banerjee, S., Saha, P., Mathew, B., vanchinathan, h.p., Mukherjee, A.: Multilingual abusive comment detection at scale for indic languages. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems, vol. 35, pp. 26176–26191. Curran Associates, Inc., ??? (2022). https://proceedings.neurips.cc/paper_files/paper/2022/file/a7c4163b33286261b24c72fd3d1707c9-Paper-Datasets_and_Benchmarks.pdf

[77] Fast, Accurate, and Scalable Video Content Moderation. ShareChat Data Science Blog (2023). https://sharechat.com/blogs/data-science/fast-accurate-and-scalable-video-content-moderation

[78] Twitch to Permanently Ban Streamers for Sexual Harassment and Assault. The Verge (2020). https://www.theverge.com/2020/6/25/21303185/twitch-sexual-harassment-assault-permanent-bans-streamers

[79] Twitch Streaming and the MeToo Reckoning: Sexual Misconduct Allegations. Wired (2020). https://www.wired.com/story/twitch-streaming-metoo-reckoning-sexual-misconduct-allegations/

[80] New Malware Spreads Over Twitch Chat, Targets Steam Accounts. PCWorld (2023). https://www.pcworld.com/article/435208/new-malware-spreads-over-twitch-chat-targets-steam-accounts.html

[81] How to Spot Phishing Scams. Uber (2023). https://www.uber.com/en-AU/blog/how-to-spot-phishing-scams/

[82] Alarming Rise of Uber Sexual Assaults. Kherkher Garcia LLP (Year Not Provided). https://www.kherkhergarcia.com/alarming-rise-of-uber-sexual-assaults/

[83] Managing Harassment on Twitch. Twitch (Year Not Provided). https://safety.twitch.tv/s/article/Managing-Harassment?language=en_US

[84] Twitch Introduces Mature Content Classification Labels for Streams. The Verge (2023). https://www.theverge.com/2023/6/20/23767736/twitch-mature-content-classification-labels-streams

[85] Combating Targeted Attacks on Twitch. Twitch (Year Not Provided). https://safety.twitch.tv/s/article/Combating-Targeted-Attacks?language=en_US

[86] Understanding Why Driver Partners Lose Account Access. Uber (Year Not Provided). https://www.uber.com/en-IN/blog/understanding-why-driver-partners-lose-account-access/

[87] Uber Driver App: Fraud Activities. Uber (Year Not Provided). https://www.uber.com/gb/en/drive/driver-app/fraud-activities/

[88] Uber Community Guidelines: Respect. Uber (Year Not Provided). https://www.uber.com/us/en/safety/uber-community-guidelines/respect/

[89] DoorDash Dasher Help: SafeChat. DoorDash (Year Not Provided). https://help.doordash.com/dashers/s/article/SafeChat?language=en_US

[90] DoorDash Unveils New Safety Features for Delivery People. TechCrunch (2022). https://techcrunch.com/2022/11/14/doordash-new-safety-features-for-delivery-people/

[91] Samdesk Partners with DoorDash to Help Keep Dashers Safe. Samdesk (Year Not Provided). https://www.samdesk.io/blog/samdesk-partners-with-doordash-to-help-keep-dashers-safe

[92] How DoorDash Ensures Velocity and Reliability Through Policy Automation. DoorDash Engineering Blog (2022). https://doordash.engineering/2022/09/20/how-doordash-ensures-velocity-and-reliability-through-policy-automation/

[93] DoorDash to Give Delivery Workers More Safety Rights and Notifications. The Verge (2022). https://www.theverge.com/2022/11/15/23460247/doordash-delivery-worker-safety-rights-notifications-block

[94] Does Bumble Have an Age Limit? Dude Hack (Year Not Provided). https://dude-hack.com/does-bumble-have-an-age-limit/

[95] Phan, A., Seigfried-Spellar, K., Choo, K.-K.R.: Threaten me softly: A review of potential dating app risks. Computers in Human Behavior Reports **3**, 100055 (2021) https://doi.org/10.1016/j.chbr.2021.100055

[96] Online Safety Dating Apps. INEQE Group (2022). https://ineqe.com/2022/02/14/dating-apps/

[97] Is Bumble Safe? Tips for Online Dating Safety. Safes.so (Year Not Provided). https://www.safes.so/blogs/is-bumble-safe/

[98] Violent Sexual Predators Use Dating Apps as Hunting Grounds for Potential Victims, Study Finds. WPDE (Year Not Provided). https://wpde.com/news/spotlight-on-america/violent-sexual-predators-use-dating-apps-as-hunting-grounds-for-potential-victims-study-finds-tinder-bumb

[99] OKCupid Security Flaws. Tom's Guide (Year Not Provided). https://www.tomsguide.com/news/okcupid-security-flaws

[100] OKCupid Security Flaw Threatens Intimate Dater Details. Threatpost (Year Not Provided). https://threatpost.com/okcupid-security-flaw-threatens-intimate-dater-details/157809/

[101] Bumble Launches Stand for Safety Initiative with Women's Safety Guide to Combat Online Abuse. Firstpost (Year Not Provided). https://www.firstpost.com/tech/news-analysis/bumble-launches-stand-for-safety-initiative-with-women-safety-guide-to-combat-online-abuse-9457981.html

[102] Safety Resources on OkCupid. OkCupid (Year Not Provided). https://help.okcupid.com/hc/en-us/articles/7952653965837-Safety-Resources

[103] Dreyfuss, E.: The Challenges of Content Moderation: Language and Artificial Intelligence. Wired (Year Not Provided). https://www.wired.com/story/content-moderation-language-artificial-intelligence/

[104] In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation. IEEE Spectrum (2016). https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation

[105] Facebook: Our AI Tools Failed to Catch New Zealand Attack Video. The Wall Street Journal (2019). https://www.wsj.com/articles/facebook-our-ai-tools-failed-to-catch-new-zealand-attack-video-11553156141

[106] Twitter's Algorithm Failure Shows Big Tech's Ongoing Struggle with AI Bias. Spiceworks (2023). https://www.spiceworks.com/tech/artificial-intelligence/news/

twitters-algorithm-failure-shows-big-techs-ongoing-struggle-with-ai-bias/amp/