Studying the Online Deepfake Community

Brian Timmerman, Pulak Mehta, Progga Deb, Kevin Gallagher, Brendan Dolan-Gavitt, Siddharth Garg and Rachel Greenstadt

Abstract. Deepfakes have become a dual-use technology with applications in the domains of art, science, and industry. However, the technology can also be leveraged maliciously in areas such as disinformation, identity fraud, and harassment. In response to the technology's dangerous potential, many deepfake creation communities have been deplatformed, including the technology's originating community: r/deepfakes. Opening in February 2018, just eight days after the removal of r/deepfakes, MrDeepFakes (MDF)—see content warning below—went online as a privately owned platform fulfilling the role of community hub, claiming to be the largest deepfake creation and discussion platform currently online. This position of community hub is contrasted against the site's main purpose of hosting nonconsensual deepfake pornography. In this paper we explore two primary deepfake communities via a mixed methods approach, applying quantitative and qualitative analysis. We identify how these platforms have been used by their members, what opinions these deepfakers hold about the technology and how it is seen by society at large, and we identify opinions regarding deepfakes-as-disinformation. We find that there is a mix of technical discussion and potentially malicious content, while the deplatforming of early deepfake communities impacted trust regarding alternative community platforms.

Content Warning

The MDF main platform is used for hosting deepfake pornography. In this paper we analyze a subsection of the platform dedicated as a discussion area, the platform's forum section.

1 Introduction

The media manipulation method known as "deepfakes"[1] is a dual-use technology that has found adoption in the domains of art, science, and industry—for example, application in VFX development pipelines and artistic self-expression (Caporusso 2020; Etienne

1. For a technical definition of deepfakes and a broader discussion of their scope, see Westerlund 2019.

doi:10.54501/jots.v2i1.126

2021; Farid 2022; Hsiang 2020; Kwok and Koh 2021; Accenture Labs 2021; Mair 2020; Neethirajan 2021); however, in addition to these applications, the technology also has the potential to inflict serious harm in application domains such as mis- and disinformation, identity fraud, and harassment. The history of deepfakes is firmly rooted in the latter category, as the community's founding hub, r/deepfakes, became widely known as a source of nonconsensual pornographic deepfake content in December 2017 during the subreddit's first month of activity. Due to the nature of the community's output, word quickly spread about the technology and its applications, catching the attention of both journalists and lawmakers (Schiff, Murphy, and Curbelo 2018; Cole 2017; McNamara 2017; Vincet 2017). This, in turn, resulted in mainstream social media platforms such as Twitter, Reddit, and Discord banning and removing many of the deepfake communities and discussions that had developed on their sites (Fingers 2018; Hern 2018; Johnson 2018).

Following this widespread deplatforming, portions of the deepfake community migrated to dedicated platforms to continue discussing deepfake technology and to share their creations, the self-proclaimed largest platform being MrDeepFakes.com (MDF). Since its founding in February 2018, just over a week after the removal of r/deepfakes, MDF has developed into a central hub for deepfake activity.[2] The primary function of MDF is the hosting of nonconsensual deepfake pornography; however, many users engage in extensive discussion on the site's forum sections regarding a wide array of deepfake-related topics. These conversation areas cover domains such as technical assistance, dataset sharing, general deepfake discussion, and an actively growing deepfake market (primarily for people seeking to commission not-safe-for-work (NSFW) deepfake media).

Deepfakes pose a serious and substantive security threat when created with malintent, a recent example being a deepfake targeting Ukrainian president Volodymyr Zelensky designed to inject disinformation in a military context (Allyn 2022; Metz 2022). Within the domain of relationships, it has been noted that deepfakes provide an attack vector to enact intimate partner abuse, violating one's likeness and engaging in reputation destruction (Lucas 2022). Additionally, a study examining the perspectives of Wikipedia editors has shown that the potential to have one's likeness artificially injected into pornographic content can create a chilling effect when engaging with online platforms (Forte, Andalibi, and Greenstadt 2017). In the context of business fraud, deepfakes have been identified as a threat enabling marketplace deception (Mustak et al. 2023). Due to the potential severity of deepfake disinformation, harassment, and fraud, it is important to understand the culture surrounding the technology, and to understand how members of this community view deepfake applications such as disinformation and fraud.

In this paper we explore the development of two key deepfake discussion platforms, r/deepfakes and MDF, and measure how these spaces have been and were utilized by their members. We employ quantitative analysis to identify platform level of activity, outline which discussion areas on these platforms areas are most active, examine to what extent members have engaged with these platforms, and quantify the growth of the emerging deepfake market.

In identifying how these platforms were and are used, we find the most common type of discussion is technical assistance, but that parallel to these technical conversations there is an actively growing deepfake market in which 98.5% of demand is for nonconsensual deepfake pornography.

2. At time of writing the site is referenced as "the biggest NSFW English community" on the GitHub page for the popular deepfaking tool DeepFaceLab (Iperov,). Additionally, MDF claims "MrDeepFakes is the largest deepfake community still actively running, and is dedicated to the members of the deepfake community." on the forum's bottom banner, and the forum has seen activity spanning over 7,600 threads.

In conversations where datasets on targets are shared on MDF, we find that the primary targets of deepfakes are celebrities, such as actors and musicians, but that there has also been attention given to potential disinformation targets such as politicians and business executives, raising concerns for potential disinformation generation.

While the r/deepfakes archived posts do not lend themselves to thematic analysis due to their fragmented nature, the threads within MDF's "Discussion" sub-forum contain nuanced and wide-ranging discussions on deepfakes and the culture surrounding them. Exploring the thoughts of deepfakers posting on MDF, we find a history of deplatforming has resulted in many perceiving MDF as their only viable option for learning and utilizing a deepfake skill set, due to perceptions of hostility from the wider culture. Through thematic analysis we identify common ethical arguments and concerns presented regarding deepfakes, explore how deepfakers attempt to navigate the legal status of their creations, and examine how deplatforming has impacted the community's trust regarding outsiders.

2 Background and Related Work

This work utilizes online community measurement methods to create a holistic profile of the past and current deepfake communities r/deepfakes and MDF. While previous work has begun to examine the cultural impact of deepfakes online, our work expands this research to cover the deepfake-dedicated community of MDF directly.

2.1 Online Community Measurement

Online community measurement (OCM) techniques have been leveraged in many areas, enabling the characterization and profiling of the groups of interest. OCM has seen usage in profiling hate speech and disinformation propagation (Hine et al. 2017; Mondal, Silva, and Benevenuto 2017), cybercriminal groups (Afroz et al. 2014; Afroz et al. 2013; Motoyama et al. 2011), communities encouraging malicious activity (Franklin et al. 2007; Tseng et al. 2020), and social media platforms in general (Cheng, Liu, and Dale 2013; Chun et al. 2008).

We apply techniques used by these studies to better understand the deepfake communities r/deepfakes and MDF, examining how users have utilized these spaces and quantifying to what depth and breadth.

2.2 Deepfake Community/Online Impact Analysis

There is a rich body of work exploring technical aspects of deepfakes, such as deepfake creation, detection, and detection circumvention (Güera and Delp 2018; Huang et al. 2020; Koopman, Rodriguez, and Geradts 2018; Li and Lyu 2018), but work surrounding the deepfake creation community is still limited. While previous work has aimed to provide a cultural context for understanding the response to deepfakes in the abstract (Burkell and Gosse 2019), our work generates such a context by directly exploring the voiced opinions of those within the MDF community. Previous works examining the deepfake creation community have explored the context of deepfake pornography (Popova 2020), the types of content that can be acquired from MDF (Kikerpill 2020), the complex legal status of deepfakes (Kugler and Pace 2021), and the ease with which novice creators can produce deepfakes (Mehta et al. 2023), and have discussed how deepfakes are emerging as a new media entity (Hsiang 2020). In the domain of community analysis, recent work has examined the nature of discourse surrounding deepfakes on Reddit from 2018 to 2021, and an interview study examined the thoughts of a specific deepfake

tool's community and their stances on deepfake ethics (Gamage et al. 2022; Widdler et al. 2022). We seek to expand on this work by examining the deepfake community on MDF, providing an analysis of how NSFW deepfake culture has changed since the deplatforming of r/deepfakes, and performing the first measurement study of MDF.

Within the domain of the impact of deepfakes on online discourse, prior work has examined areas such as the responses of YouTube users regarding high view count deepfake videos, as well as a deep dive analysis on a specific deepfake of actor Keanu Reeves and the originating YouTube channel (Lee et al. 2021; Bode 2021). We expand upon this work by examining deepfake tutorial videos rather than deepfakes themselves, exploring how the tools referenced in these videos may act as a gateway to the platform of MDF.

3 Communities and Dataset

Due to the emergent nature of deepfake technology, and as a result of community take-downs, few deepfake discussion platforms have had the ability to form and fully develop. In our search for active deepfake communities other than MDF, we identified a number of Discord groups and forums dedicated to specific deepfake creation tools, a subreddit for SFW deepfakes, and a Telegram group used to discuss the tool DeepFaceLab. However, at time of data collection the Discord servers had only existed for a limited number of months and lacked the range of detailed conversation found on MDF. Furthermore, the Telegram group, while having a considerable amount of activity, is primarily used by the Russian deepfake community, preventing analysis by our team.

A number of readily available learning resources have potential to introduce deepfakers to MDF. We sampled thirty deepfake tutorials on YouTube (found by searching "deepfake tutorial"), and identified 56.7% of videos either directly linked to the DeepFaceLab[3] or DeepFaceLive[4] GitHub repositories. The DeepFaceLab repository links to MDF eight times, advertising it as a space for technical assistance and community discussion, while the DeepFaceLive repository links to it once as a communication group. The videos containing links or references to these tools accounted for 86.6% of total views across all videos, accumulating 4,198,042 views total at time of writing.[5] This direction to find help on MDF is compounded by a lack of technical assistance on other mainstream platforms. When searching "deepfacelab" on Stack Overflow at time of writing, only fifteen user questions were returned, of which only six had received answers. These findings are in line with MDF's own claim to be "the largest deepfake community still actively running," and accordingly we consider it to be a center of the current deepfake community environment.

While we consider r/deepfakes and MDF to be primary deepfake communities, we explore each for different reasons. In the case of r/deepfakes, this was the originating community that was used to facilitate deepfake discussion during the technology's initial release. To the best of our knowledge, no other deepfake discussion space was meaningfully utilized during the time this subreddit was online. In the case of MDF, we have not been able to identify any other community with the scope of discussion areas related to deepfakes and the quantity of users on this platform. While the primary use of the platform is the hosting of deepfake pornography, the forums section we examine encompasses discussion regarding every facet of deepfakes; we will discuss these topic

3. https://github.com/iperov/DeepFaceLab
4. https://github.com/iperov/DeepFaceLive
5. Due to the personalized and temporal nature of YouTube video recommendations, the specific videos and view counts at time of sampling have been recorded in Appendix B.

Table 1: Discussion Categories

Category Name	Sub-forums Within Category
General Discussion	Discussion
Technical Resources and Assistance	Guides and Tutorials, Tools and Apps, Questions
Creation Resource Sharing	Unofficial Mods, Trained Models, Pornstar Facesets, Celebrity Facesets, Celebrity to Pornstar Matches, Celebrity Faceset Requests, Downloads
Content Sharing	Celebrity Deepfakes, Celebrity Photo Fakes, SFW Deepfake Videos
Content Requests	Requests (NSFW), Requests (SFW), Requests (Image Deepfakes)
Meta Discussion	Announcements and News, Claim Credit/Flag Videos, How To Use Site Features, Suggestion and Feedback

areas further in Section 4.2. We note that this analysis only considers spaces dedicated to deepfake discussion specifically, not considering deepfake discussion on general web platforms, but due to the funneling of users to MDF via technical resources and its rooting in r/deepfakes, we have chosen to limit the scope of our research to focus on these platforms.

The total data collected from MDF spans over 5.2K threads, 20K posts, and almost 4K accounts from February 2018 to August 2021, while the data from r/deepfakes consists of 1.2K threads, 4.6k comments, and 1.5K accounts from December 2017 to February 2018. Data from MDF was collected via a custom scraper, while data from r/deepfakes was collected from the Pushshift API (Pushshift,). Before moving to our analysis, we discuss the format of MDF and elaborate on the nature of our data and collection methods.

3.1 MrDeepFakes Forum Structure

MDF has a number of sub-forum sections dedicated to different aspects of deepfake creation and services. We group these sub-forums into unique topic categories based on common purpose, and later use them to identify what discussion areas are most popular among MDF users. The construction of these categories is found in Table 1, and will be discussed in depth in Section 4.2.[6]

3.2 Data Specifications

Data from MDF was collected using a custom-built web scraper written in Python, leveraging the BeautifulSoup web-browsing package. Our script strictly collected only text content hosted on the forums, ignoring all other forms of media. Upon completing collection, all usernames were scrubbed and replaced with unique identifiers, and during manual analysis of data all personal information contained in viewed posts was removed. All data collected is publicly available without need of an MDF site account to view. The time period captured in the data spans from the site's creation in February 2018 to

6. A number of the sub-forums are dedicated to *facesets*. A faceset is a collection of face images, normally extracted from videos, used as training data during deepfake creation.

August 2021, for a total of over 26,000 posts. The data was collected on a per-post basis, maintaining information regarding:

- Author identifier and forum rank/role
- Author's total thread and post counts
- Post date and time of creation
- Post text content, post likes, and any authors the post is responding to
- Title of thread being responded to

To collect data from the banned r/deepfakes, we utilized the Pushshift API to gather archived threads and comments from the subreddit (Pushshift,). This archival data covers a time range from mid-December 2017 to early February 2018 for original thread posts, and from mid-December 2017 to mid-January 2018 for comments on those threads. We note that the scope of the archival data is incomplete, with comments from the final month missing from our collection. We also note that a number of posts do not include the author's username, due to being marked as "[deleted]." In the data collected, we found 157 instances of deleted account usernames. Additionally, the amount of threads and posts that were removed from r/deepfakes before archival could be conducted is unknown. We utilize the following data returned by the Pushshift API for each original thread post:

- Author's Reddit account name
- Date of the post
- Content of the post
- Title of the thread
- Total upvotes given to thread
- Number of comments left on thread

The subreddit's unstructured nature results in topics across all discussion domains being combined in one pool, requiring topic labeling on a thread-by-thread basis. Manual labeling was deemed required due to the extremely broad range of discussion areas, which resulted in low coherency outputs from automated analysis methods such as LDA topic modeling. To generate these labels, a member of our team manually parsed through the collected threads, identifying common areas of discussion and labeling the thread data according to identified topic areas. Additional details and the results of this labeling are discussed further in Section 4.2.

3.2.1 Ethics and Data Management

During our custom data collection and analysis of the MDF forums, we worked closely with our IRB to ensure all guidelines were followed, and we ensured steps were taken to minimize the storage of sensitive information. Data was stored on a secure, institutional hosting platform, usernames were replaced with unique identifier codes, and personal information was removed whenever found in posted text.

4 Platform Usage Dynamics

To examine how these deepfake communities developed over time, we generate and analyze platform usage statistics from r/deepfakes and the MDF forums. In this section

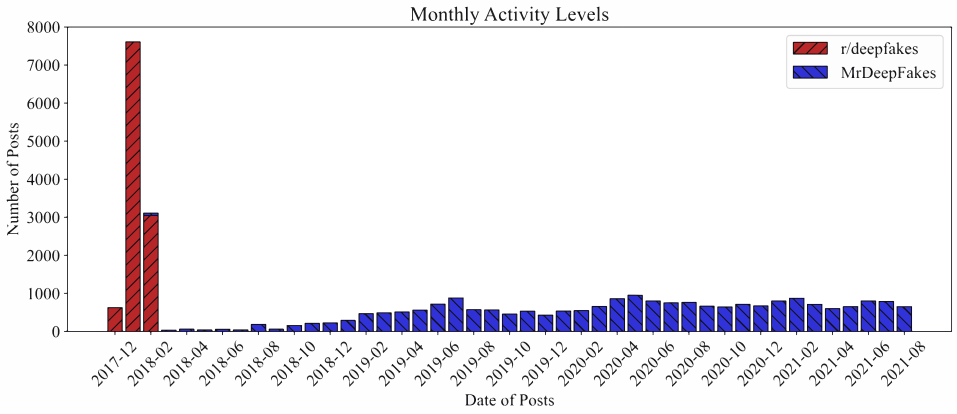

Figure 1: Monthly Activity Levels: r/deepfakes and MrDeepFakes

we examine total usage levels over time, enumerate common conversation domains to measure their comparative levels of activity, and explore to what extent users, on average, interact with these communities in terms of breadth and depth.

4.1 Platform Activity Levels

The first measurement we present is overall usage levels describing r/deepfakes and MDF activity, cataloging the total amount of posts made per month. In total, r/deepfakes existed for approximately two months (our data for r/deepfakes spans from December 14, 2017, to February 7, 2018) before it was banned from Reddit. Our data for MDF starts just over a week after the subreddit's removal, with the oldest post occurring on February 15, 2018, covering a time range through August 2021. As discussed in Section 3, the time span for archived comments on r/deepfakes posts is incomplete; however, the Pushshift API returns the number of comments per thread at time of archival, which we use to determine activity levels such that our results are minimally impacted by the period missing in the archive. The levels of monthly activity identified on these platforms are presented in Figure 1.

After establishing a baseline of user activity in 2018–2019, from the beginning of 2020 through to the end of August 2021 MDF averaged 724.5 posts per month, approximately 9.5% of the monthly activity seen during the peak of r/deepfakes in January 2018. Our data shows that MDF has stabilized around this activity level, with a monthly usage minimum of 550 posts and a maximum of 953 posts during this time span.

4.2 Conversation Topic Categories

Users of r/deepfakes and MDF have discussed a wide range of topics regarding deepfakes, such as technical questions, data/content sharing, and issues of deepfake ethics and legality. We measure to what extent these deepfake communities discuss common topic areas, and examine how much engagement conversations in these domains receive.

Due to differences in platform structure, different measurement approaches need to be applied when considering r/deepfakes threads versus MDF content. As discussed in Section 3 and displayed in Table 1, MDF is well structured such that we can easily identify what topic a thread covers via which sub-forum it was submitted to, whereas r/deepfakes hosted all topics together in the same discussion pool. To properly measure the scope of conversation on r/deepfakes, we manually parsed through the initial posts in each

thread, identifying common conversation domains and labeling them according to topic area. Unfortunately, the content of many of the archived posts was not captured, instead displaying as either "[deleted]" or "[removed]." However, in many these cases we are able to use the thread's title to determine original intent when explicitly clear (Example title: "Any Smart Dudes: Please do some Selena Gomez, serious lack of her here!"). Instances where the title was vague and the message deleted were discarded (Example title: "A few tests"). Of the 1,256 r/deepfakes threads we collected, we were able to determine thread topic in 1,127 instances, resulting in 90% coverage. To note, this labeling and identification of discussed topics is meant only to introduce what general conversations areas posters engaged with on r/deepfakes, rather than draw any deeper insight into the themes of those discussions. The common topic areas found on r/deepfakes during our analysis were the same as what was found on MDF, consisting of:

- **General Discussion.** Discussions about the technology in general, including discussions on deepfake ethics, legality, censorship/community suppression, and alternative application areas
- **Technical Resources and Assistance.** Discussions relating to errors encountered making deepfakes, issues occurring with the creation tools themselves, and suggestions for improvements to those creation tools
- **Creation Resource Sharing.** Users sharing datasets used to create deepfakes, sharing resources that can be used to generate data for a model, and sourcing high-quality videos for data ripping
- **Content Sharing.** Users sharing their deepfakes, and users requesting someone re-post a previously deleted deepfake
- **Content Requests.** Discussions surrounding who should be deepfaked, who a target should be swapped onto for a high-quality output, and deepfake requests for specific targets
- **Meta Discussion.** Discussions about the platform itself, rules, milestones, and announcements

The category usage results for r/deepfakes are contained in Table 2, with MDF's results in Table 3.

Table 2: r/deepfakes: Category Usage Statistics

Category	Threads	Comments	Avg Thread Length	Avg Thread Upvotes
General Discussion	143 (11.4%)	1,989 (19.8%)	13.9	25.0
Technical Resources and Assistance	618 (49.2%)	5,120 (51.0%)	8.3	3.6
Creation Resource Sharing	40 (3.2%)	251 (2.5%)	6.2	15.0
Content Sharing	53 (4.2%)	467 (4.7%)	8.8	21.6
Content Requests	221 (17.6%)	913 (9.1%)	4.1	3.9
Meta Discussion	52 (4.1%)	872 (8.7%)	16.8	41.1
Unknown Topic	129 (10.3%)	418 (4.2%)	3.2	6.3

We find the dominating discussion category in both r/deepfakes and MDF is technical discussions, comprising 49.2% of all threads in our r/deepfakes data and 42% of all threads on MDF. Upon manual review of the technical discussion threads hosted on

Table 3: MDF: Category Usage Statistics

Category	Threads	Comments	Users	Avg Thread Length	Avg Thread Views
General Discussion	233 (4.5%)	826 (4%)	328 (6.2%)	4.54	2,125
Technical Resources and Assistance	2188 (42%)	12,582 (60.2%)	1,961 (36.9%)	6.75	2,670
Creation Resource Sharing	1,069 (20.5%)	3,007 (14.4%)	1108 (20.9%)	3.81	1,989
Content Sharing	805 (15.4%)	1,680 (8.0%)	765 (14.4%)	3.09	1,887
Content Requests	742 (14.2%)	2,216 (10.6%)	974 (18.3%)	3.99	1,108
Meta Discussion	182 (3.5%)	591 (2.8%)	173 (3.3%)	4.25	1,240

MDF, we found six overarching discussion areas: deepfake software, hardware, general assistance, data processing, model training, and post processing.

Where the two platforms deviate is in activity per thread, measured by thread length. On MDF, threads contained in the Technical Discussion category see the highest levels of activity per thread, versus r/deepfake's General and Meta Discussion domains. Within threads on r/deepfakes that we labeled as belonging to General Discussion, we identified that the most active threads pertained to ethics (31.1 comments and 36.3 upvotes on average over 27 threads) and legality (14.1 comments and 18.2 upvotes on average over 13 threads).

4.3 User Engagement

In order to understand user engagement, we measure how broadly users engaged with these platforms (how many different categories users post in) and how frequently (how many posts posters make on average). We find both platforms' users typically engage/engaged with only one aspect of deepfake creation, with 77.8% and 93.6% of posters on MDF and r/deepfakes, respectively, only commenting in one topic domain, indicating shallow engagement by most platform users.[7]

When examining the number of posts made per poster, usage statistics show the overwhelming majority of users of both MDF and r/deepfakes are casual, with 92.7% and 95.0% of the user populations, respectively, only making between 1 to 10 posts in total. Combined with the topic breadth statistics, we find most individuals who post on these platforms are limited in their post count, and tend to use the platforms for only one deepfake-related topic area (with the largest topic area being Technical Discussion).

We note the overall usage patterns of both r/deepfakes and MDF are similar, in terms of both discussion areas and user engagement. When r/deepfakes was banned en masse, removing both content sharing/requests as well as technical assistance at the same time, these community aspects had to be reestablished elsewhere. This is the dynamic of the dual nature of the MDF platform: developing to provide a marketplace for pornographic deepfakes while also acting as a hub of technical knowledge. This combination of platform functions as a product of holistic community removal serves as a cautionary tale of the effects of blanket deplatforming within developing technology communities.

7. Due to the incomplete nature of the archived r/deepfakes data, we are only able to measure this metric for original thread posters on r/deepfakes, whereas we are able to explore all post authors on MDF.

Having explored overall platform usage levels, what topic domains attract the most attention, and how users tend to engage with these platforms, we next consider the potential threat of disinformation, community isolation as a defense mechanism, and the nature of the deepfake market on MDF.

5 Detailed MDF Sub-forum Examination

Having performed high level measurements on both r/deepfakes and MDF, we now focus our attention on three specific MDF sub-forums. We specifically examine MDF content, rather than r/deepfakes, due to multiple factors. As r/deepfakes only existed for a short duration (approximately two months), the content only reflects a very small snapshot in time. Additionally, by focusing on MDF, we are able to focus on and identify post-community deplatforming opinions and dynamics. In the upcoming sections we consider the MDF community's preferred deepfake targets, how the community perceives outside entities with a lack of trust, and deepfake market dynamics.

5.1 "Celebrity Facesets" Sub-forum: Target Profiles

To make a deepfake, creators require a substantial amount of training data, called "facesets" for their target models. This can be a time-intensive process, requiring manual face masking for high-quality training output. To address this, many high-profile individuals have had premade facesets published on the "Celebrity Facesets" sub-forum, enabling users to quickly acquire training data and begin deepfaking these targets. Reviewing this data, we found 455 unique individuals discussed, and in this section we examine how that faceset data is distributed over individuals of different professions, nationalities, and genders. We utilize the number of views the threads covering these individuals receive, and use this view count data to gauge interest in these targets by the community.[8] The measurements for target gender, occupation, and nationality are contained in Tables 4 and 5. All individuals listed on this sub-forum are public figures such as politicians, celebrities, social media personalities, business leaders, journalists, and athletes. All information was collected via public individual profiling sites, online biographical listings, and the public social media accounts of named individuals. In instances where a target was listed as having dual citizenship, the count for both countries was incremented.

We find that there is a 2:1 ratio of female to male deepfake targets, and also the same ratio in the number of views they receive. In terms of occupations, we found 77.7% of targets came from an entertainment background (Celebrities, Internet Personalities, Athletes, and Voice Actors), while 22.3% of targets have clear potential for disinformation production (Political Figures, Business Executives, Journalists/Newscasters, Religious Figures). When pooling by these categories, we find targets with major risk of usage in disinformation receive 720 views on average, versus 1,246 for entertainers and celebrities. This presents a 1.73:1 ratio between interest in standard public celebrities and potential disinformation targets. When profiled by nationality, we found targets belonging to 73 nations, Table 5 contains the results for nations that have more than five deepfake targets on the sub-forum.[9] We find the majority of targets come from English-speaking countries such as the United States, the United Kingdom, Canada, and Australia, but that there are significant presences from India and South Korea.

Taken as a whole, these statistics indicate the most common target for deepfaking within

8. When an individual is covered in multiple threads, we take the view count of the most viewed thread, aiming to avoid potential confounding factors that may arise from combining thread statistics.

9. For a full list of countries, see Appendix C on page 28.

Table 4: Deepfake Targets by Gender

Gender	Total Count	Avg. Views/Thread
Female	313 (68.8%)	1,361
Male	142 (31.2%)	614

Table 5: Deepfake Targets by Nationality

Nationality	Total Count	Avg. Views/Thread
American	229 (50.3%)	1,160
British	39 (8.6%)	1,617
Indian	39 (8.6%)	1,481
Canadian	28 (6.2%)	1,085
South Korean	13 (2.9%)	1,424
Australian	12 (2.6%)	1,458
Russian	7 (1.5%)	999

this community is a female celebrity from a native English-speaking country. However, there are also a significant number of individuals listed on this sub-forum with potential for disinformation application, but figures of this type are not posted as frequently nor do they attract as much attention.

Having examined what targets are popular for deepfaking, we now examine MDF's "Discussion" sub-forum, exploring what deepfakers think on topics such as ethics, legality, and the direction of the technology.

5.2 'Discussion' Sub-forum: Thematic Analysis

To capture the thoughts and opinions of active MDF posters, we conducted thematic analysis of 148 threads from the platform's "Discussion" sub-forum. This sub-forum contains discussions on a wide range of topics, ranging from the ethics of deepfakes to legal concerns, and provides insight into a diverse spectrum of perspectives.

5.2.1 Thematic Analysis: Reasoning and Methodology

During our research multiple analysis types were considered when approaching the question of MDF user sentiment and opinions. We first attempted to utilize automated LDA topic modeling methods to gain insight into these areas, but this approach generated low-coherency topic bundles due to the large diversity of conversation areas. We also advertised paid interviews for deepfake producers; however, despite the advertisement thread receiving over 250 views, only two individuals responded to our outreach directly and only one sat for an interview. Accordingly, manual review of thread contents through thematic analysis was deemed the appropriate way to approach our research question.

Thematic analysis has seen use in a number of online measurement studies for the purpose of opinion and sentiment identification originating from online forums (Attard and Coulson 2012; Moore, Ayers, Drey, et al. 2016; Smedley and Coulson 2017; Chivers et al. 2020). To generate our codebook we followed the process for inductive, data-driven code creation (DeCuir-Gunby, Marshall, and McCulloch 2011), consisting of:

- **Reducing the data:** We split our set of 148 threads among four coders on our team, such that each coder had a unique set of threads. These coders then began to analyze their assigned threads while paying attention to relations that arose between posts.
- **Identification of themes within sub-samples:** As the coders explored their threads, themes that appeared during the back-and-forth of the forum members participating in the thread's conversation were identified.
- **Comparing themes across sub-samples:** Coders considered common themes and sentiments that regularly occurred across conversations.
- **Code creation:** Upon review and analysis of the data, the team members generated codes that covered the themes identified within their thread subsets.
- **Reliability and overlap:** After each member had generated their codes, the team reconvened to discuss their results. A discussion round occurred where a finalized set of codes was agreed upon. The finalized codes were produced with the objective of encapsulating all themes identified during the process, while merging similar codes into generalized topics.

We note that our final code set was determined utilizing coder agreement (CA) rather than via inter-rater reliability (IRR). As discussed in "Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice" (McDonald, Schoenebeck, and Forte 2019), both CA and IRR have advantages and disadvantages when producing a final codebook. We determined CA was appropriate for our investigation as it allows for synthesizing emergent themes via discussion of individual coder's readings of their data subset.

Upon creation of our finalized codebook, a second round of coding was then conducted by a single primary coder, using the group-made codebook, on the full set of "Discussion" threads. After this final coding was conducted, the team again reconvened to double-check the finalized coding results and to discuss their meaning. We now present the overarching themes and perspectives identified during this coding process; the finalized codebook is contained in Appendix A for reference.

5.2.2 Ethics and Legality

Posters frequently engaged in contentious dialogue regarding the ethics of deepfake creation, the legality of what they are doing, and what values the community has as a whole. The following sections provide examples of some of the common opinions forum members voiced on these topics.

Ethics

Many conversations discussing the ethical considerations of deepfake creation appeared in our data. Interestingly, a vocal population of forum posters believe that pornographic deepfake creation without consent is unethical. When explaining their stances, members who have ethical concerns cite the lack of consent involved, and assert that using someone's likeness in a situation they do not approve of cannot be morally right. However, while a portion of MDF's posters believe nonconsensual deepfake pornography is ethically dubious, others present varied justifications for why this content is acceptable. The community's most repeated lines of reasoning fall into the following categories[10]:

10. All quotations from MDF posts are presented as they appeared on the platform.

- Deepfakes do not cause direct harm: *"Violence is an act committed against a person that results in bodily harm, physical abuse is violence for example, sexual assault is violence, a video is not violence. Even a fake video of fake violence is not violence."*
- Deepfakes are a new art form in an early stage: *"But even clay modelling and cave paintings was first used for porn. There is always resistance to novelty. But even if there might have been a temporary ban on those Arts, they still went on. Same here. We just gotta show some diversity."*
- Deepfakes are fundamentally the same as preexisting media editing tools: *"Technically, this is a cropping of one face onto another body. Has been done since decades. But for pictures. Now the pictures are moving."*
- They themselves are comfortable being deepfaked: *"I'd be thrilled that someone found me attractive enough to be worth making a deepfake of lol!"*
- NSFW content is what is in demand: *"I make porn because thats what people want.... I actually don't find the porn that arousing, I am just in it for the tech and learning."*

While we were unable to find any mention of disinformation explicitly, we were able to find discussions regarding deepfakes being used in cases of fraud or deception. In all conversations surrounding these use cases, members stated they were against such applications, and in some cases, supported limiting the technology due to its capacity for harm. The reasoning for these stances included ethical concerns as well as theorizing such applications would increase legal restrictions. In our analysis we found no discussion of deepfakes for disinformation or fraud in a positive context.

Legality

Legal concerns are another frequent discussion topic on MDF. Due to the emergent nature of deepfake technology, limited legal precedent has been set regarding acceptable applications, manifesting in posters seeking reassurance they won't be punished for their deepfake content. Some users cite feelings of paranoia, and others voice concerns they may be breaking the law by watching deepfakes at all.

> *"Hey all. I make a fuckton of deepfakes and intend to continue doing so. I'd love to start posting more to this site. But, I'm tremendously paranoid about the dangers of doing so."*

> *"Can a person who broadcasts or watches Deepfake videos be penalized? What can the courts or the police do in this situation? So what would your reaction be if he was arrested? I wonder about these, my friends"*

> *"Good questions regarding the legality around this. I must say there is currently no clear answer but making deepfakes is technically NOT illegal. I have tried to make certain rules regarding which deepfakes will be hosted on this site, but of course as this tech attracts more users, people can use it to do harm (portray deepfakes as real, defamation, etc) and host the videos elsewhere."*

Looking forward to the future of deepfake regulation, the overall tone of members turns fatalistic, as posters believe it's only a matter of time before their works become outlawed. However, there is a counter line of reasoning—some posters believe that, due to the open source nature of modern deepfake tools, laws won't do much to prevent deepfake creation, and as long as there are active developers the deepfake scene will always exist.

> *"Even if the legislation is not there yet to make this illegal, it will come."*

> *"I don't think it's a silly question, I worry about the same thing! With the rise of revenge porn laws and the deepfake technology improving I do feel it may be only a matter of time before this stuff is deemed illegal"*

> *"Deepfakes will not die. This is now opensource technology and there are active developers. And as long as the tech is available, people will always enjoy porn"*

Along with the concerns posters have about the legality of their activity, users discuss the nature of content and labor ownership in deepfakes. Because deepfakes are a method of modifying preexisting content, creators often use copyrighted content during the creation process. Ownership discussions focus on the right of deepfake creators to profit from their creations and the morality of others stealing their deepfakes. This creates a tension between creators who believe they own their deepfakes and should be able to profit from them (citing the transformative labor being applied), and those who believe deepfakes don't have a legal justification of ownership (citing the impact of deepfaking is not sufficiently transformative).

> *"So just got another Copyright Strike on my xvideos account although I heavily edited the deepfake (deleted audio, edited color exposure)....They are completely ignoring fair use agreements..."*

> *"Welcome to the modern internet where everything is copyrighted, people copyright strike people on youtube for stuff they don't own, where monetization is taken from you"*

> *"Deepfakes are already on the fringe of legality to begin with, and the ones that use copyrighted material and charge for them certainly are not legal. Bottom line, there's really no good way to keep people from stealing and reposting."*

> *"Well the face of the celebrity has been used without permission & I would assume that the porn being used was not home made. It is like a rap song that takes music from one place & the lyrics from another. None of these are an original work. I understand that there has been time committed to put these together, and I appreciate it, but this isn't a Mona Lisa forgery"*

5.2.3 Community Suppression and Outside Trust

As the MDF forums were created during an era of mass deepfake community deplatforming, concerns are frequently voiced about what could happen to MDF and the posters that use it. The following quote comes from early in MDF's existence in August 2018:

> *"After the devastating bannwave from reditt, gfycat, twitter, discorde and "pornhub" the community lost a lot of its members. And now with the closing of dpfak.com it happened again i'm afraid the a0adult deepfake scene won't survive anothe blow like that."*

A separate exchange showed members having difficulty accessing the site due to an ISP-level block:

> *"Since I moved I've been unable to access this site on my home wifi. I'm accessing right now using mobile data and that works just fine. When I run a*

> *traceroute to this domain I get an IP belonging to ddos-guard.com I've spent hours on the phone with my ISP and they insist they aren't blocking anything. Is it possible the MrDF webserver is blocking traffic from my IP? Quick update: I just tried from behind a VPN and it works. Someone is clearly blocking something here. Any similar experiences?"*
>
> *"I had no issues a few months ago, but now I cant access this site without using a vpn."*
>
> *"Hey all, thanks for the replies. I confirmed I can access the site while behind a VPN, and after spending hours and hours in calls / chats with* <name of ISP> *they did in fact admit that they 'block some sites for security reasons' but thats as far as i got with them. I guess ill add a vpn to my monthly budget for now!"*

This awareness of the communities' fragility, in combination with the negative attention received from the media, has created a privacy-focused culture wary of outsiders and alternative community hosting platforms. It's presumed that communities for deepfakes hosted outside of MDF will eventually be deplatformed due to public hostility toward deepfake technology. Additionally, some members are concerned about the future of MDF itself, concerned outside forces may force the site down. Some users reported issues accessing the site in June 2021 due to an ISP block, heightening these concerns.

> *"I noticed a lack of a discord for DeepFaceLab so I put one together real quick."*
>
> *"That's surprising, usually then ban this stuff, similar to reddit, etc."*
>
> *"The discord is great, I've learned more on the deepfake discord for this site (there's another, slow one that's SFW) than I have on the site almost. But we'll get banned eventually"*

Seeing MDF as the last safe and stable platform to discuss deepfakes without risk of being banned, the community often shares advice on how members can protect themselves as deepfake creators. Members give privacy and plausible deniability advice to concerned community members regarding areas such as:

- Encrypting data related to deepfake works
- Proper VPN usage
- Using burner accounts and false information
- Adding disclaimer watermarks to their deepfakes to prevent claims of disinformation

This defensive stance further manifests in the community's opinions toward media outlets covering deepfake stories. Some members feel strong animosity directed at the media, as they see them as a driving force pushing deepfakes off major platforms through unfair coverage. In instances in which media organizations have reached out to interview members of the forums, community leaders have urged caution.

> *"As for other uses of deepfakes, there are plenty of SFW content on YouTube or reddit. The media just doesn't cover that as much as the porn."*
>
> *"I think there should be some concern. As deepfakes get more media attention, porn websites and video hosts will be pressured in fighting against hosting*

> *these videos. They may disrupt the deepfake scene, but I don't think it'll go away that easily."*
>
> *"For those thinking about contacting the media to discuss deepfakes, please remember to protect your identity. Regardless if you make SFW or NSFW deepfakes, there is always a negative stigma around this topic."*

However, some members see the media coverage as an opportunity to grow the community via press coverage, even suggesting MDF should actively attempt to garner attention. These members contend that is because of the media, rather than despite it, that the deepfake creation community has grown as it has.

> *"I get that some of you might be afraid of anothe media shittstorm, but it was BECAUSE of the media 100.000 people tuned in."*
>
> *"i cant accept that this is it, we need to do a massive pr campaign on facebook, twitter, youtube and reddit to make adult deepfakes what it was before !"*

On the whole, the MDF community has a complex relationship with outside groups and platforms. Many assume outside entities will be hostile by default, due to the widespread deplatforming seen when the technology appeared and the widely negative press coverage. At best, outside forces are seen as a means to an end to build the community via advertisement. Even within MDF there is a culture of threat mitigation, as members believe they could face retribution due to public sentiment surrounding their creations.

5.2.4 Monetary Gain

Along with casual deepfake discussions, some members use the forums as a platform to discuss business opportunities arising out of the growing deepfake ecosystem. Enterprising forum members have discovered potential markets such as:

- Selling commissioned deepfakes to private buyers
- Donations from supportive content viewers
- Charging "tokens" that can be exchanged for money as payment for video downloads
- Providing technical aid and assistance
- Custom deepfake creation resources (such as facesets and custom trained models)
- Selling web URLs related to deepfakes

> *"Hi, we are looking for deepfake artists who can make quality content from non-celebrity. We are a creative online marketing agency in* <nationality> *where we set up campaigns from A to Z. We have recently been receiving a lot of requests from customers interested in deepfake videos for their campaigns. As we see the popularity of these videos increasing, we are looking for a company that specializes in making these videos. If someone is interested, they can email their portfolio and price to* [contact email]*"*

5.2.5 Direction of Deepfake Technology and Use Cases

Another area of contention within the MDF community is the direction of deepfake creation technology and its potential use cases. Due to the explicitly pornographic nature of the site, many users feel that should be the focus of community creations, while other members, wanting to expand into other domains, feel stifled. The following exchange encapsulates this dynamic:

> *"Bruh not all are making deep fake porn. Some are making normal deep fakes like some scenes in movies etc where they would love to not just see the fake but HEAR it too in its original voice."*
>
> *"You are on a porn site dude."*
>
> *"DeepFake is not JUST porn. I am on Special porn/non porn DeepFake site"*
>
> *"Yes most here are here for the deepfake porn (making). But it is commonly known that "war" and "porn" is the evolutional drive for progress and development. So this is the place to learn and get the knowledge you need to become a successful deepfake creator. Not everybody is a smut peddler."*

From the members that have discussed alternative application areas, the following have been speculated:

- Interactive VR Scenes
- Movie Special Effects
- Educational Content
- Artistic Creation
- Fantasy Fulfillment (such as deepfaking yourself into movie scenes)

Others are beginning to consider the place deepfakes have in wider development pipelines, such as VFX pipelines.

> *"Hi I'm a really experienced VFX artist working mostly on TV Commercials and some movies. I do a lot of face and beauty work so this new avenue of deepfakes is really interesting for me. I'm always interested in new ground breaking projects and ideas. If you have an amazing project I may well be interested in touching it up to make it more realistic."*

5.3 Content Requests: The Deepfake Market

The MDF market sub-forum is home to a growing deepfake marketplace comprising three primary areas: NSFW Videos, SFW Videos, and NSFW Images. Each of these three areas has "Paid" and "Free" solicitation areas. The usage levels for these categories are contained in Table 6. We note that 98.5% of all threads requesting deepfake content are in the NSFW domain.

This is a notable change from what was seen on r/deepfakes, which had a negligible number of paid solicitations. We found that of the twenty paid creation/services posts found on r/deepfakes, only fourteen were a direct solicitation for a paid deepfake commission, while the rest were concerned with technical assistance and domain names. Across the approximate two-month lifespan of r/deepfakes, this results in an average of seven paid requests per month. Comparatively, during the month of August 2021,

Table 6: Content Request Sub-forum Activity on MDF

Sub-forum Purpose	Threads	Comments	Posters
NSFW Video: Paid	448	822	521
NSFW Video: Free	271	1,322	511
SFW Video: Paid	4	7	9
SFW Video: Free	7	46	20
NSFW Image: Paid	1	1	2
NSFW Image: Free	11	18	13

the most recent month of MDF posts in our dataset, there were a total of 70 threads on the paid request sub-forum, a 900% increase in average monthly activity from the paid commission discussions on r/deepfakes. We find the self-governed and well-structured nature of MDF has enabled the community to create a formalized location for a NSFW deepfake market, which has facilitated an elevated level of paid commissions for pornographic deepfakes.

The scope of these markets are difficult to gauge due to a "direct message" culture. On the previously mentioned "Paid Request" sub-forum, the average number of replies is only 1.8 per thread, whereas the average number of views is 442.5. This discrepancy between views and posts results in uncertainty regarding the number of active deepfake producers working for profit. Of the messages left on request threads, the average response is a form of "direct message me for more info." This locks casual observers out of viewing price negotiations, the going rate for deepfake commissions, and the activity level of the market.

6 Discussion

In this work we have presented the first measurement study of the prolific deepfake community MrDeepFakes. We have tracked community activity levels on two primary deepfake discussion platforms, examined the impact deplatforming r/deepfakes had on MDF's community growth and poster outlook, and analyzed the MDF community's stances on disinformation and trust regarding outside entities. We have also enumerated what common topics of conversation occur within these spaces, and identified the average profile of deepfake targets on these platforms.

Through a combination of qualitative and quantitative analysis, we find community usage dynamics to be complex. While a majority of discussions within both MDF and r/deepfakes were technical in nature, and many members claimed they want to remain lawful and ethical, the primary content type produced or demanded within these spaces has been of nonconsensual, pornographic deepfakes. In addition, it must be noted there is a presence of facesets that do present concerns along mis- and disinformation lines.

Examining the discussion of members on why they use MDF suggests that this tension may be due to a lack of trust in alternative platforms. A wariness of treatment in mainstream spaces appears to lend to an environment where members feel MDF is the only viable platform for learning a deepfake skill set, even if said users intend to leverage the skills in socially accepted positions such as VFX studios. This has created a space where novice deepfakers are introduced to a market that primarily demands nonconsensual deepfakes of pornographic content.

When enumerating the individuals listed on the Celebrity Faceset sub-forum, we have

identified the most common deepfake target profile is women from native English-speaking countries who work in entertainment. This may be due in part to the previously discussed market demand for deepfakes of such targets. However, we also find there are a significant number of targets with listings on this sub-forum who pose serious risk for disinformation creation, with politicians, business leaders, religious figures, and news anchors comprising 22.3% of all listings.

As discussed in Section 3, there is little in the way of deepfake resources on traditional technical assistance platforms such as Stack Overflow, on which we only discovered 15 mentions of the technology. Regarding why MDF maintains its popularity as the primary deepfake discussion platform, as compared to mainstream alternatives, our thematic analysis indicates the community lacks trust when engaging with outside entities. As one user wrote regarding a deepfake Discord server:

> *"It's gonna get banned probably, even if it's SFW."*

From discussions of mitigating deplatforming, fears regarding ISP blocks, and the lack of outreach in response to our interview offers, it appears the deepfake community on MDF has insulated as a defense mechanism.

While the ethics of deepfake applications are under intense discussion, it is undeniable that roles producing deepfakes with monetary incentive, such as VFX artists, are increasingly finding use cases for the technology. We believe that alternative spaces dedicated to learning deepfakes should be constructed, so that those learning to make deepfakes for socially accepted roles with monetary incentives can do so in a well-regulated domain. Such a space would enable learning a deepfake skill set without exposure to a market primarily demanding pornographic, nonconsensual content. For such a platform to succeed, buy-in from industry professionals would be required to compile technical knowledge as well as provide mentorship and expertise.

The story of the r/deepfakes and MDF communities contains lessons and warnings applicable to AI-tool communities at large. As further AI generation methods receive general adoption by technology enthusiasts, such as the ChatGPT large language model and stable diffusion art generators, we will see additional burgeoning communities focused on these tools. If and when these technologies are used to generate unethical or dangerous content, those who have power over the platforms used to discuss these methods must take care in their moderation and management. As we have shown in this work, blanket actions to suppress discussion may result in the opposite of the intended outcome, amplifying the undesired application areas in newly formed dedicated spaces.

7 Future Work

During our analysis we found that a considerable amount of activity on MDF consists of technical discussions that would be acceptable on platforms such as r/SFWdeepfakes. Additional work should examine why MDF has been and continues to be the preferred hub of technical assistance for deepfakes compared to mainstream platforms. Furthermore, work should explore the population that renders technical assistance on MDF, and examine what motivating factors encourage them to do so.

While this work provides measurement of many aspects of the deepfake community, there are areas that we were unable to explore. As discussed in Section 5.3 there is a complex, emerging market surrounding deepfakes that largely exists away from public eyes. Future work should explore the scope and costs of deepfake services that are

currently available online, enumerating what individuals with no deepfake experience can buy from skilled producers. Additionally, while this work spans public community posts from r/deepfakes to MDF, it strictly examines publicly made posts. While our efforts to reach out to community members did not receive much engagement, additional efforts should be attempted to facilitate dialogue.

8 Conclusion

In this paper we have analyzed the two primary online deepfake communities, MrDeepFakes and r/deepfakes. Utilizing both qualitative and quantitative analysis methods, our profiles show complex platform usage dynamics. On both platforms we found the most common content type is purely technical in nature, but it exists alongside a large demand for nonconsensual, pornographic deepfake material. Looking to the thoughts and opinions voiced by posters on MrDeepFakes, we found a culture of distrust toward alternative spaces regarding deepfake discussion, as well as a defensive posture regarding how deepfakes are seen by society as a whole. This has contributed to an environment where the most resource-rich place to learn a deepfake skill set is also the primary market hub for nonconsensual deepfake pornography.

References

Accenture Labs. 2021. *Flipping the script on deepfake technologies,* September. https://www.accenture.com/tw-en/insights/technology/deepfake-technologies.

Afroz, Sadia, Vaibhav Garg, Damon McCoy, and Rachel Greenstadt. 2013. "Honor among thieves: A common's analysis of cybercrime economies." In *2013 APWG eCrime Researchers Summit,* 1–11. IEEE. https://doi.org/10.1109/ecrs.2013.6805778.

Afroz, Sadia, Aylin Caliskan Islam, Ariel Stolerman, Rachel Greenstadt, and Damon McCoy. 2014. "Doppelgänger finder: Taking stylometry to the underground." In *2014 IEEE Symposium on Security and Privacy,* 212–26. IEEE. https://doi.org/10.1109/sp.2014.21.

Allyn, Bobby. 2022. "Deepfake video of Zelenskyy could be 'tip of the iceberg' in info war, experts warn." *NPR* (March). https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia.

Attard, Angelica, and Neil S Coulson. 2012. "A thematic analysis of patient communication in Parkinson's disease online support group discussion forums." *Computers in Human Behavior* 28 (2): 500–506. https://doi.org/10.1016/j.chb.2011.10.022.

Bode, Lisa. 2021. "Deepfaking Keanu: YouTube deepfakes, platform visual effects, and the complexity of reception." *Convergence* 27 (4): 919–34. https://doi.org/10.1177/13548565211030454.

Burkell, Jacquelyn, and Chandell Gosse. 2019. "Nothing new here: Emphasizing the social and cultural context of deepfakes." *First Monday,* https://doi.org/10.5210/fm.v24i12.10287.

Caporusso, Nicholas. 2020. "Deepfakes for the good: A beneficial application of contentious artificial intelligence technology." In *International Conference on Applied Human Factors and Ergonomics,* 235–41. Springer. https://doi.org/10.1007/978-3-030-51328-3_33.

Cheng, Xu, Jiangchuan Liu, and Cameron Dale. 2013. "Understanding the characteristics of internet short video sharing: A youtube-based measurement study." *IEEE Transactions on Multimedia* 15 (5): 1184–94. https://doi.org/10.1109/tmm.2013.2265531.

Chivers, Bonnie R, Rhonda M Garad, Jacqueline A Boyle, Helen Skouteris, Helena J Teede, and Cheryce L Harrison. 2020. "Perinatal distress during COVID-19: thematic analysis of an online parenting forum." *Journal of Medical Internet Research* 22 (9): e22002. https://doi.org/10.2196/22002.

Chun, Hyunwoo, Haewoon Kwak, Young-Ho Eom, Yong-Yeol Ahn, Sue Moon, and Hawoong Jeong. 2008. "Comparison of online social relations in volume vs interaction: a case study of cyworld." In *Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement,* 57–70. https://doi.org/10.1145/1452520.1452528.

Cole, Samantha. 2017. "AI-Assisted Fake Porn Is Here and We're All Fucked." *Motherboard* (December 11, 2017). https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn.

DeCuir-Gunby, Jessica T, Patricia L Marshall, and Allison W McCulloch. 2011. "Developing and using a codebook for the analysis of interview data: An example from a professional development research project." *Field Methods* 23 (2): 136–55. https://doi.org/10.1177/1525822x10388468.

Etienne, Hubert. 2021. "The future of online trust (and why Deepfake is advancing it)." *AI and Ethics,* 1–10. https://doi.org/10.1007/s43681-021-00072-1.

Farid, Hany. 2022. "Creating, Using, Misusing, and Detecting Deep Fakes." *Journal of Online Trust and Safety* 1 (4). https://doi.org/10.54501/jots.v1i4.56.

Fingers, Jon. 2018. "Reddit bans the 'deepfake' AI porn it helped spawn." *Engadget* (February). https://www.engadget.com/2018-02-07-reddit-bans-deepfake-ai-porn.html.

Forte, Andrea, Nazanin Andalibi, and Rachel Greenstadt. 2017. "Privacy, Anonymity, and Perceived Risk in Open Collaboration: A Study of Tor Users and Wikipedians." In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing,* 1800–1811. CSCW '17. Portland, Oregon, USA: Association for Computing Machinery. ISBN: 9781450343350. https://doi.org/10.1145/2998181.2998273.

Franklin, Jason, Adrian Perrig, Vern Paxson, and Stefan Savage. 2007. "An inquiry into the nature and causes of the wealth of internet miscreants." *Ccs* 7:375–88. https://doi.org/doi.org/10.1145/1315245.1315292.

Gamage, Dilrukshi, Piyush Ghasiya, Vamshi Bonagiri, Mark E Whiting, and Kazutoshi Sasahara. 2022. "Are Deepfakes Concerning? Analyzing Conversations of Deepfakes on Reddit and Exploring Societal Implications." In *CHI Conference on Human Factors in Computing Systems,* 1–19. https://doi.org/10.1145/3491102.3517446.

Güera, David, and Edward J. Delp. 2018. "Deepfake Video Detection Using Recurrent Neural Networks." In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS),* 1–6. https://doi.org/10.1109/AVSS.2018.8639163.

Hern, Alex. 2018. "'Deepfake' face-swap porn videos banned by Pornhub and Twitter." *The Guardian* (February 7, 2018). https://www.theguardian.com/technology/2018/feb/07/twitter-pornhub-ban-deepfake-ai-face-swap-porn-videos-celebrities-gfycat-reddit.

Hine, Gabriel, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. 2017. "Kek, Cucks, and God Emperor Trump: A measurement study of 4chan's politically incorrect forum and its effects on the web." In *Proceedings of the International AAAI Conference on Web and Social Media,* vol. 11. 1. https://doi.org/10.1609/icwsm.v11i1.14893.

Hsiang, Emily. 2020. *Deepfake: An Emerging New Media Object in the Age of Online Content.* Student paper. https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=9014787&fileOId=9014794.

Huang, Yihao, Felix Juefei-Xu, Run Wang, Qing Guo, Lei Ma, Xiaofei Xie, Jianwen Li, Weikai Miao, Yang Liu, and Geguang Pu. 2020. "Fakepolisher: Making deepfakes more detection-evasive by shallow reconstruction." In *Proceedings of the 28th ACM International Conference on Multimedia,* 1217–26. https://doi.org/10.1145/3394171.3413732.

Iperov. *DeepFaceLab.* GitHub repository. Accessed August 16, 2023. https://github.com/iperov/DeepFaceLab.

Johnson, Khari. 2018. "AI Weekly: Can tech platforms police themselves in a deepfake-filled future?" *VentureBeat* (February 9, 2018). https://venturebeat.com/2018/02/09/ai-weekly-can-tech-platforms-police-themselves-in-a-deepfake-filled-future.

Kikerpill, Kristjan. 2020. "Choose your stars and studs: The rise of deepfake designer porn." *Porn Studies* 7 (4): 352–56. https://doi.org/10.1080/23268743.2020.1765851.

Koopman, Marissa, Andrea Macarulla Rodriguez, and Zeno Geradts. 2018. "Detection of deepfake video manipulation." In *The 20th Irish Machine Vision and Image Processing Conference (IMVIP),* 133–36. ISBN: 978-0-9934207-3-3. http://hdl.handle.net/2262/89508.

Kugler, Matthew B, and Carly Pace. 2021. "Deepfake Privacy: Attitudes and Regulation." *Northwestern Public Law Research Paper,* nos. 21-04, https://doi.org/10.2139/ssrn.3781968.

Kwok, Andrei O. J., and Sharon G. M. Koh. 2021. "Deepfake: a social construction of technology perspective." *Current Issues in Tourism* 24 (13): 1798–802. https://doi.org/10.1080/13683500.2020.1738357.

Lee, YoungAh, Kuo-Ting Huang, Robin Blom, Rebecca Schriner, and Carl A Ciccarelli. 2021. "To believe or not to believe: framing analysis of content and audience response of top 10 deepfake videos on YouTube." *Cyberpsychology, Behavior, and Social Networking* 24 (3): 153–58. https://doi.org/10.1089/cyber.2020.0176.

Li, Yuezun, and Siwei Lyu. 2018. "Exposing deepfake videos by detecting face warping artifacts." *arXiv preprint arXiv:1811.00656,* https://doi.org/10.48550/arXiv.1811.00656.

Lucas, Kweilin T. 2022. "Deepfakes and domestic violence: perpetrating intimate partner abuse using video technology." *Victims & Offenders* 17 (5): 647–59. https://doi.org/10.1080/15564886.2022.2036656.

Mair, Chazz. 2020. "New Deepfake Tech Lets Companies Use You In Commercials." *Screenrant* (January 8, 2020). https://screenrant.com/deepfake-tech-company-commercials/.

McDonald, Nora, Sarita Schoenebeck, and Andrea Forte. 2019. "Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice." *Proceedings of the ACM on Human-Computer Interaction* 3 (CSCW): 1–23. https://doi.org/10.1145/3359174.

McNamara, Brittney. 2017. "AI Porn Raises Issues of Consent." *Teen Vogue* (December 11, 2017). https://www.teenvogue.com/story/ai-porn-consent.

Mehta, Pulak, Gauri Jagatap, Kevin Gallagher, Brian Timmerman, Progga Deb, Siddharth Garg, Rachel Greenstadt, and Brendan Dolan-Gavitt. 2023. "Can deepfakes be created on a whim?" In *Companion Proceedings of the ACM Web Conference 2023,* 1324–34. https://doi.org/10.1145/3543873.3587581.

Metz, Rachel. 2022. "Facebook and YouTube say they removed Zelensky deepfake." *CNN* (March 16, 2022). https://www.cnn.com/2022/03/16/tech/deepfake-zelensky-facebook-meta/index.html.

Mondal, Mainack, Leandro Araújo Silva, and Fabrício Benevenuto. 2017. "A measurement study of hate speech in social media." In *Proceedings of the 28th ACM Conference on Hypertext and Social Media,* 85–94. https://doi.org/10.1145/3078714.3078723.

Moore, Donna, Susan Ayers, Nicholas Drey, et al. 2016. "A thematic analysis of stigma and disclosure for perinatal depression on an online forum." *JMIR Mental Health* 3 (2): e5611. https://doi.org/10.2196/mental.5611.

Motoyama, Marti, Damon McCoy, Kirill Levchenko, Stefan Savage, and Geoffrey M Voelker. 2011. "An analysis of underground forums." In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference,* 71–80. https://doi.org/10.1145/2068816.2068824.

Mustak, Mekhail, Joni Salminen, Matti Mäntymäki, Arafat Rahman, and Yogesh K Dwivedi. 2023. "Deepfakes: Deceptions, mitigations, and opportunities." *Journal of Business Research* 154:113368. https://doi.org/10.1016/j.jbusres.2022.113368.

Neethirajan, Suresh. 2021. *Beyond Deepfake Technology Fear: On its Positive Uses for Livestock Farming.* Preprint. https://doi.org/10.20944/preprints202107.0326.v1.

Popova, Milena. 2020. "Reading out of context: pornographic deepfakes, celebrity and intimacy." *Porn Studies* 7 (4): 367–81. https://doi.org/10.1080/23268743.2019.1675090.

Pushshift. *Pushshift Reddit API Documentation.* GitHub repository. Accessed August 16, 2023. https://github.com/pushshift/api.

Schiff, Adam B., Stephanie Murphy, and Carlos Curbelo. 2018. *Letter to Daniel R. Coates.* Congressional Letter, September. https://schiff.house.gov/imo/media/doc/2018-09%5C%20ODNI%5C%20Deep%5C%20Fakes%5C%20letter.pdf.

Smedley, Richard M, and Neil S Coulson. 2017. "A thematic analysis of messages posted by moderators within health-related asynchronous online support forums." *Patient Education and Counseling* 100 (9): 1688–93. https://doi.org/10.1016/j.pec.2017.04.008.

Tseng, Emily, Rosanna Bellini, Nora McDonald, Matan Danos, Rachel Greenstadt, Damon McCoy, Nicola Dell, and Thomas Ristenpart. 2020. "The tools and tactics used in intimate partner surveillance: An analysis of online infidelity forums." In *29th USENIX Security Symposium (USENIX Security 20),* 1893–909. ISBN: 978-1-939133-17-5. https://www.usenix.org/conference/usenixsecurity20/presentation/tseng.

Vincet, James. 2017. "AI tools will make it easy to create fake porn of just about anybody." *The Verge* (December 12, 2017). https://www.theverge.com/2017/12/12/16766596/ai-fake-porn-celebrities-machine-learning.

Westerlund, Mika. 2019. "The emergence of deepfake technology: A review." *Technology Innovation Management Review* 9 (11). https://doi.org/10.22215/timreview/1282.

Widdler, David Gray, Dawn Nafus, Laura Dabbish, and James Herbsleb. 2022. "Limits and Possibilities for "Ethical AI" in Open Source: A Study of Deepfakes." In *2022 ACM Conference on Fairness, Accountability, and Transparency.* https://doi.org/10.1145/3531146.3533779.

Authors

Brian Timmerman is a PhD candidate at New York University.

(brian.timmerman@nyu.edu)

Pulak Mehta was a Masters student at New York University at the time of this research.

Progga Deb was a Masters student at New York University at the time of this research.

Kevin Gallagher is an Assistant Professor at the NOVA School of Science and Technology and an integrated member at NOVA LINCS.

Brendan Dolan-Gavitt is an Associate Professor at New York University.

Siddharth Garg is an Institute Associate Professor at New York University.

Rachel Greenstadt is an Associate Professor at New York University.

Acknowledgements

We thank the MDF users and moderators who responded to our requests and questions.

Data availability statement

The r/deepfakes data can be acquired via the Pushshift API. The posts collected from MDF are not available due to containing potential personally identifying information.

Funding statement

This work is supported by NOVA LINCS (UIDB/04516/2020) with the financial support of FCT.IP and by National Science Foundation (NSF) Award 2016061.

Ethical standards

The post data collection portion of this study was deemed exempt by the NYU IRB as part of IRB-FY2021-5561. The interview portion (results not included due to not enough participants) was approved by the NYU IRB as part of IRB-FY2022-6290.

Keywords

deepfakes; disinformation; trust; deplatforming

Appendices

Appendix A: Thematic Analysis Codebook

The following codes were used to analyze the “Discussion” sub-forum data in Section 5.2:

- Ethics, Morals, Values — Personal / Community
- Legality
- Community Suppression
- Suppression Circumvention
- Monetary Gain
- Intra-community Assistance
- Creation Motivation
- Deepfake Technology Use Cases
- Direction of Deepfake Technology
- Labor and Content Ownership

Appendix B: YouTube Deepfake Tutorials

Table 7 lists the deepfake tutorials sampled on YouTube, including the sample's URL, view count at time of sampling, and whether the video is tangentially related to MDF.

Table 7: YouTube Tutorial Samples

Video Link	MDF Tangential	View Count
https://www.youtube.com/watch?v=t59gRbpYMiY	Yes	1,585,910
https://www.youtube.com/watch?v=lSM-9RBk3HQ	Yes	786,179
https://www.youtube.com/watch?v=Zmutd9618Kk	Yes	620,899
https://www.youtube.com/watch?v=wYSmp-nrJ7M	Yes	277,330
https://www.youtube.com/watch?v=hP3njGbvvWc	Yes	196,051
https://www.youtube.com/watch?v=tW7EENTWXRk	Yes	195,374
https://www.youtube.com/watch?v=0p-nNSvB7KA	Yes	117,565
https://www.youtube.com/watch?v=q44LPygdMxU	Yes	97,360
https://www.youtube.com/watch?v=QSmHho1uHFM	Yes	61,504
https://www.youtube.com/watch?v=_bc3SPbCdW8	Yes	50,173
https://www.youtube.com/watch?v=rw5F5lSvBLE	Yes	46,926
https://www.youtube.com/watch?v=CHs3VuW7TtU	Yes	38,157
https://www.youtube.com/watch?v=ljMXS8vovx4	Yes	37,301
https://www.youtube.com/watch?v=N6XN0fjHZSA	Yes	36,514
https://www.youtube.com/watch?v=1Bt5wyGqdk4	Yes	32,432
https://www.youtube.com/watch?v=zpOclEB2-dk	Yes	15,845
https://www.youtube.com/watch?v=k0Z-ZuQ1fi8	Yes	2,522
https://www.youtube.com/watch?v=qaqRLopz0wA	No	309,593
https://www.youtube.com/watch?v=bQihTE3QBVk	No	87,235
https://www.youtube.com/watch?v=pud_1PUoFXA	No	78,782
https://www.youtube.com/watch?v=is347MG71yY	No	50,703
https://www.youtube.com/watch?v=OI1LEN-SgLM	No	50,016
https://www.youtube.com/watch?v=VBNCusXEhA8	No	25,945
https://www.youtube.com/watch?v=5CU81pJAPV4	No	16,364
https://www.youtube.com/watch?v=W8xaFW-2UR8	No	8,602
https://www.youtube.com/watch?v=ev1rhuBzBCw	No	7,782
https://www.youtube.com/watch?v=rw0oavibpF8	No	6,736
https://www.youtube.com/watch?v=KLe7-3bMpg4	No	3,131
https://www.youtube.com/watch?v=MjyBswm3dG4	No	2,410
https://www.youtube.com/watch?v=DQGRD9KfdCw	No	436

Appendix C: Full List of Deepfake Target Nationalities

The following is a full list of the countries to which individuals discussed in the "Celebrity Faceset" sub-forum of MDF belong:

- America
- Argentina
- Australia
- Austria
- Belarus
- Belgium
- Bolivia
- Brazil
- Britain
- Bulgaria
- Canada
- Chile
- China
- Columbia
- Croatia
- Cuba
- Cyprus
- Czech Republic
- Denmark
- Ecuador
- Estonia
- Finland
- France
- Georgia
- Germany
- Greece
- Guatemala
- Guyana
- Hungary
- Iceland
- India
- Ireland
- Israel
- Italy
- Japan
- Latvia
- Lithuania
- Luxembourg
- Macedonia
- Malta
- Mexico
- Moldova
- Monaco
- Montenegro
- Morocco
- Netherlands
- New Zealand
- Norway
- Pakistan
- Paraguay
- Peru
- Philippines
- Poland
- Portugal
- Romania
- Russia
- Scotland
- Serbia
- Slovakia
- Slovenia
- South Africa
- South Korea
- Spain
- Suriname
- Sweden
- Switzerland
- Thailand
- Turkey
- Ukraine
- Uruguay
- Venezuela