ANALYZING SOCIETAL AND ETHICAL CONCERNS OF WEAPONIZING SOCIAL MEDIA

CLYDE VILLACRUSIS, ALEXANDER CHEN, ETHAN NAHLINDER, OSCAR BASUYAUX

1. Executive Summary

Social media has been widely used by many people, ranging from advertising to content creation. It has several benefits such as improving lives through monetization and sharing useful information. However, it has many drawbacks. The primary drawback is the process of weaponization in social media, especially as it pertains to shaping public opinion and hurting lives. Firstly, people take advantage of bots as a means to troll other people, as well as posting misinformation. This has gained attention similar to the 2016 US presidential election where the Russians leveraged Facebook and Twitter to intervene with the election process. This is evidence that social media is prone to vulnerabilities and is thus crucial to mitigate these damages for the sake of our democracy and the wellbeing of society. In addition, ever-growing cyberbullying takes place in various forms in social media. People perform cyberbullying through Instagram comments — whether it is for a joke or not —, or through text messages where perpetrators can get away easily. Moreover, social media algorithms utilize the data of each user to personify their feeds, making the user feel satisfied with their content. And perpetrators can easily use this tactic to spread dangerous information to other people, like the Israel vs. Hamas conflict. For the ISIS case, their managers used social media to instill fear and cause panic by using hashtags like #ALLEYESONISIS, which also plays a crucial role in their propaganda.

There are 3 primary problems on why it is difficult to deal with the weaponization of social media: complex detection of fake accounts and automated bots, sophisticated modern disinformation campaigns, and limitations of detection algorithms and AI. Detecting fake accounts and bots is a challenge because bots are programmed to mimic human-like actions. As for fake accounts, it can be deactivated for a long time then reactivated in time of an important event that can cause people to sway from one information perspective to another. In addition, sophisticated modern disinformation campaigns can employ abusive tactics and online harassment. Lastly, the limits of machine learning (ML) and detection algorithms causes an issue to real-world problems because it cannot fully harness the visual image techniques. Visual imaging techniques are crucial to detecting cyberbullying and harsh comments as the majority of people can get away with texts.

Cyberbullying mostly occurs to individuals who have less power than the perpetrators. The main place cyberbullying occurs is through peer-to-peer interactions over texts. In a study of university students in a psychology course, 31% of cyberbullying victims said that they did nothing when they were being cyberbullied, while 25% of them reported. Moreover, the majority of the participants reported

that they use text conversations. Another similar study was conducted on Reddit, a social media platform. As a result of this study, many perpetrators felt comfortable doing the cyberbullying aggressions because they are behind the screens all the time. Moreover, they were less likely to be aggressive towards peers they know. To combat cyberbullying, many researchers utilized machine learning techniques, such as sentiment analysis where they analyze each comment to see if it is a positive, neutral, or negative stance. However, there are many problems that occur with cyberbullying detections, such as imbalance between text and images, and a shared definition of harsh words by a machine learning model. For the imbalance, it is currently hard for ML to detect what is actually happening, e.g., something actually happened in TikTok, but ML detected it on Instagram. Therefore, ML scientists will need to improve their algorithms and ensure that detecting visual images are more accurate and detect cyberbullying instances from one social media to another.

Moving forward to the Israel-Palestine conflict, the way they handle their misinformation for their propaganda is stated as follows. Israel-Palestine has been through conflicts for the past 70 years. In addition, there is the infamous group called Israeli Defense Forces (IDF). They used social media to voice their narrative, thus controlling people. They also employed a sophisticated and well-funded approach, leveraging these platforms to share infographics, videos, and live updates on military operations. For instance, they shared videos of airstrikes, and posted infographics detailing their military successes and the threats they faced from Hamas . This strategy is not just about communication but also psychological warfare, aiming to demoralize opponents and garner international support by portraying Israel as a victim of terrorism and a defender of democratic values, especially as Israel is the only democracy in the Middle East and a strategic ally for Western powers. YourFavoriteGuy, a Tiktoker, was paid 5,000 to spread misinformation about Gaza and this is to show how Israel leveraged social media for their own benefits. Moreover, both sides of the conflict use these platforms to their benefit by mobilizing support through content that draws attention to specific political narratives.

As for ISIS and Iraq's invasion, ISIS uses social media to create posts that mainly deal with spreading their ideology through many forms of propaganda. And then they broadcast this misinformation through easily accessible channels whose vulnerable minds can absorb this imagery and thus are encouraged to join them. It has helped at least 30,000 foreign fighters from 100 countries draw to the battlefields of Syria. In addition, they have also created a domino effect in which they use social media to announce their military operations. This helps fuel the sense of the Islamic State's momentum through an attempt of "humanizing their actions", meaning they are seen visiting wounded patients in hospitals to justify their extremist mission, while still instilling fears to others through disturbing videos. So, their primary goal is to make sure the images and videos are seen at all. To combat this, many social media platforms have banned at least 10,000 ISIS-supporting accounts in a day, however, new accounts keep popping up and posting the same

information, so it is difficult to identify resurgent accounts from a large amount of data.

As for the 2016 US Presidential election, Russian hackers hacked into the email inboxes of the DNC and Hilary Clinton's campaigns, obtaining thousands of documents. One of the primary goals of Russia is to make Trump win, so they used all sorts of their social media tactics to convince the US citizens to vote for Trump. Secondly, Russia's goal was to also provoke the media such that The IRA and their team engaged in informational warfare, leading a campaign that included online political posts, targeted ads, and even political rallies inside the United States. Lastly, they also used a method called "off-the-shelf", a social media analytical software to separate users into specific demographic and interest groups. These groups would include LGBTQ+, Blue Lives Matter, and more. Once they segmented their audience into several groups, they could release targeted content with conflicting messages to each individual group.

After the showcase of four case studies about weaponization of social media, an atrocious impact on democracy can be (and has been) made, as well as the political process. Firstly, news outlets using foreign actors were able to utilize the preexisting social media platforms to spread propaganda and misinformation. They generated news articles and social media posts directed towards specific target demographics in order to misinform voters and cause chaos. Therefore, we can show how an ethical framework such as deontology, especially the reciprocity principle, apply. Applying this principle to this case, we can see how social media can be an unethical tool that makes others violate the citizens' ability to conduct a proper election. The actions of those who use social media to spread false information treat their tricked audience as means to promote their political goals and not as individuals with a right to undisturbed voting.

In addition, the use of social media in a manner to spread fake news and attack democratic systems also has meaningful social consequences. In the case of the Russian trolls of the IRA, they created accounts on both sides of online discussions and campaigns in order to create a greater social divide. Thus, a sense of social trust is somewhat jeopardized, and that sense is important in democratic nations.

Next, when considering the machine learning algorithm and how people use it for platform consideration, it can raise censorship problems. If ML algorithms are not foolproof, then it is possible that they could delete legitimate news and hide valid, political discussion. It's also possible that something could be widely accused of being fake news and end up censored, when in reality it had legitimacy. This can violate the idea of utilitarianism because if social media managers can mistakenly remove a new story headline, then it would not only limit the information available to the public, but it violates the first amendment right to free speech.

There are two recommendations to combat the weaponization of social media: improve media literacy and ML algorithms. Improving media literacy will make

individuals make more informed decisions and trust more in the media. A study was conducted showing 26.5% improvement in ability to discern between mainstream and false news headlines. In addition, there are a variety of ML algorithms, like TensorFlow and Random Forest Classification, that can further improve classification of hate comments. These algorithms go through a mathematical vector so that the model can learn and predict what the future comments are going to say.

2. Introduction

The subject matter of this report regards the weaponization of social media, specifically as it relates to shaping public opinion. It aims to answer questions regarding the use of trolls and bots in influencing user opinions on the large-scale, most importantly in the political arena. Using case studies and instances of weaponization, the report builds out a socioethical analysis of the impact of social media and, in turn, provides ethics- and technology-centered recommendations for addressing the identification and restriction of this information's large-scale dissemination.

Methodology:

In writing this report, the team has chosen to split the research topic "Weaponizing Social Media" into four distinct case studies encompassing technological, societal, and ethical problems. These considerations are also discussed at length in relation to their respective case study, as well as in the broader context of the topic and in the general ethical analysis section of the report. The case studies — Cyberbullying, Israel-Palestine, ISIS in the Invasion of Iraq, and the 2016 US Presidential Election — all touch on different aspects of the weaponization of social media, including domestic and foreign politics, religion, and societal dynamics, all relating back to social media's influence on public opinion. Each topic has been carefully chosen to expose the shortcomings of social media technology in controlling the flow and nature of sensitive and controversial content shown to millions of users daily.

3. Problem Statement

Social media platforms have revolutionized the way that humans interact and communicate, allowing for the rapid dissemination of information globally, as well as increased connectivity across numerous demographics across the world. However, in recent years, these platforms have been the subject of weaponization by various political agents looking to influence public opinion, manipulate political outcomes, and sow chaos and discord.

Part of this trend is the prevalent use of trolls and bots to post and spread misinformation, amplify controversial content, and disrupt various political processes (Barsotti, 2018; Davies, 2018). While the objective of influencing public opinion via media is a long-standing practice, the specific use of social media has garnered a lot of attention following the 2016 US presidential election, whereby Russian entities have leveraged platforms such as Facebook and Twitter to interfere with the election process (Barsotti, 2018; Davies, 2018). This incident shed light on the

vulnerabilities of social media and their serious consequences (Barsotti, 2018). It is therefore imperative to assess and address these limitations in an effort to mitigate the damage caused by ongoing and future misinformation campaigns; a large part of this issue is addressing the power and influence held by trolls and bots in political contexts.

The influence of these harmful agents extends beyond online harassment. It usually manifests via sophisticated disinformation campaigns designed to sway public opinion and undermine the establishment of governments, ideologies, and political systems (Oxford University, 2021). The use of bots – automated accounts – and trolls – coordinated human operatives – serves to effectively spread misinformation rapidly, often outpacing any effort to debunk it (Stricot, 2017). Social media platforms such as Facebook, Twitter, Instagram, or TikTok are looked upon by many as powerful tools for positive change, enabling political engagement by marginalized communities, among other things (Barsotti, 2018; Stricot, 2017). However, these platforms are also exploited for negative purposes both domestically and internationally, as in the case of the 2016 Election, the January 6th Capitol Riots, or even the Israel-Palestine Conflict. This dual nature requires a thorough understanding of social media's historical context and a comprehensive approach to regulation and intervention.

Mechanisms of Social Media Weaponization: Trolls are individuals or groups who deliberately provoke and offend people online to elicit emotional responses (Barsotti, 2018). They often post inflammatory, extraneous, or off-topic messages in online communities. Trolls can be state-sponsored or independent actors who aim to disrupt conversations and spread disinformation (Barsotti, 2018; Stricot, 2017). Bots: Bots are automated accounts programmed to perform specific tasks on social media, such as posting content, liking posts, or following users. Bots can operate at a scale beyond human capacity, amplifying messages and creating the illusion of widespread support or opposition. Advanced bots can mimic human behavior, making them difficult to detect (Barsotti, 2018; Oxford University, 2021).

Methods of social media manipulation:

These campaigns involve the deliberate creation and dissemination of false or misleading information with the goal of deceiving the public, influencing opinions, and disrupting political processes (Barsotti, 2018; Stricot, 2017).

Data-driven targeting: Social media platforms collect vast amounts of user data, which can be exploited to target individuals with tailored disinformation. By analyzing user behavior, preferences, and demographics, manipulators can craft highly effective messages that will target specific audiences (Stricot, 2017).

Echo chambers and filter bubbles: Social media algorithms create echo chambers and filter bubbles by showing users content that aligns with their existing beliefs (Barsotti, 2018). This reinforces users' viewpoints and isolates them from

opposing perspectives. Manipulators exploit this by flooding these spaces with disinformation, further polarizing the audience (Barsotti, 2018; Stricot, 2017).

Abusive strategies and harassment: Tactics such as online harassment, doxxing, and smear campaigns are used to intimidate and silence opponents. These methods can deter individuals from participating in online discussions or expressing dissenting views (Barsotti, 2018; Oxford University, 2021).

Historical Context: Social media platforms, Facebook notably, initially served as spaces for personal interaction and community building. They evolved into various use cases, including for the mobilization of grassroots campaigns and the facilitation of civic engagement. Nevertheless, their potential for community building and large-scale interpersonal interactions eventually started becoming subject to exploitation, including for spreading political messages and influencing public opinion (Oxford University, 2021). Early uses of social media in this context include the Arab Spring uprisings, which was initially marked by optimism about the democratizing potential of these tools, but quickly exposed their shortcomings in safeguarding the integrity of information and truth (Stricot, 2017).

Among poignant examples of such manipulations are the 2016 US Presidential Election and the ISIS-led invasion of Mosul. In the case of the former, Russian operatives effectively utilized Facebook and Twitter to interfere with the U.S. presidential election by spreading misinformation about candidates and parties, and creating fake accounts to disseminate this information (Stricot, 2017). The Internet Research Agency in St. Petersburg was a key player, with its activities aimed at deepening political divides and influencing voter behavior (Barsotti, 2018; Stricot, 2017). In the case of the latter, during the invasion of Mosul, ISIS employed social media strategically to broadcast their military actions, instill fear, and attract global attention. This included using hashtags like #AllEyesOnISIS to ensure their propaganda reached a wide audience, which played a crucial role in their rapid territorial gains (Davies, 2018).

Real-World Consequences: A host of real-world consequences ensued in the various cases of social media misuse. In the example of the 2016 US Presidential Election, the spread of misinformation is thought to have altered the election outcome and influence voter behavior. Research conducted by Carnegie Mellon University's Heinz College suggests that millions of Americans were exposed to such information in the build-up of the election, thereby unfairly swaying voters in a certain direction and causing many to question the integrity of the government and voting system in the aftermath (Barsotti, 2018). Similarly, in the case of the January 6th Capitol Riots, social media platforms were used to mobilize Trump supporters and other dissidents in violently breaching the US Capitol Building, further exposing the dangers of social media and the lax government policies regarding content regulation and filtration on these platforms (Stricot, 2017). These real-world consequences translated into an eroded trust in media and institutions. People have grown more skeptical of traditional news sources, increasingly resort-

ing to social media for news sources. To understand the impact of these trends, it is important to acknowledge that, cumulatively, Facebook and Twitter have 2.656 billion monthly active users, including 253.35 million users in the United States (Stricot, 2017).

Challenges in Identifying and Restricting Misinformation: The challenge in identifying and restricting misinformation across social media platforms is a three-pronged problem — (i) Complex detection of fake accounts and automated bots, (ii) Sophisticated modern disinformation campaigns, and (iii) limitations of detection algorithms and AI.

The detection of fake accounts and bots is a complex task because of modern disinformation techniques. Bots can be programmed to mimic human behavior, making them difficult to identify. Additionally, fake accounts often appear dormant for long periods before being activated for specific campaigns or agendas (Barsotti, 2018; Durso, 2023). In addition, disinformation campaigns are increasingly sophisticated, often using data-driven targeting to reach specific audiences and employ abusive tactics like smear campaigns and online harassment (Oxford University, 2021; Durso, 2023). Finally, while social media platforms have implemented algorithms and AI to detect misinformation, these tools are not failsafes (Durso, 2023). The large volume of content and the evolving nature of disinformation tactics make it challenging to effectively police platforms, even as it is an automated process by artificial intelligence and specifically-designed algorithms (Stricot, 2017; Durso, 2023).

4. Technological Issues

CYBERBULLYING

Cyberbullying is a form of virtual aggression towards someone to cause harm and it can be done in multiple ways such as mass-emailing to hundreds of people. It primarily happens to individuals that have less power over them. The main place where cyberbullying occurs is through peer-to-peer interactions by text-messaging, Instagram, and Facebook messages. However, Pyzalski, a researcher who established “6 categories” in the umbrella of cyberbullying: cyber aggression against peers, vulnerable, random cyber aggression such as victim is anonymous, cyber aggression against racial or ethnic groups, cyber against celebrities, and cyber against school staff (Pyzalski 2011). Moreover, cyberbullying mostly occurs in college campuses, although some individuals say that they have been cyberbullying since middle school (Kowalski and Whittaker 2014). In contrast, social media is the technology that people have been using for decades to advertise their business products, text their social friends, and find like minded communities and people. However, there needs to be a secure process that should be done to combat cyberbullying so that more people can be more comfortable with using social media platforms.

CASE STUDIES: In a study conducted by Kowalski and Whittaker, a total of 169 female and 75 male undergraduates taking an intro to psychology course participated where 84% are white and 8.6% are african-american (Kowalski and Whittaker 2014). They were asked which of the popular social media, i.e., Facebook, Instagram, cyberbullying mostly occur and then were asked how they reacted to victimization, i.e., did nothing, stopped the perpetrator, or reported the cyberbullying. As a result, the majority of participants used texting as their primary use of communication, followed by email and Facebook. 18.2% of participants said they have been the victim of cyberbullying at least once throughout the year, and 12% said that they initiated the cyberbullying at least once throughout the year. In addition, most of the perpetrators were either friends or a student. As for cyberbullying victimization, 25% did nothing, followed by 31.8% stating that they reported the cyberbullying.

This study suggests that with technology changing every day, there are more ways for people to do cyberbullying. The cyberbullying techniques for adolescents might be different from seniors in college and above. For adolescents, the most common technique to respond to cyberbullying is "to do nothing" ((Kowalski and Whittaker 2014), while for college students, their response is to block the person on social media platforms and report them.

Another similar study was conducted, with similar procedures. The only difference is that this one was conducted on the social media platform called Reddit. The participants reported that they were the most aggressive to people that they do not know personally, thus, they are more comfortable behind the screens. Moreover, they were less likely to be aggressive towards peers that they know personally on subreddits.

Last study conducted by the same researchers utilized the Radian6 program to extract data from different social media platforms to analyze the sentiments, as well as contexts of the comments. Then they created a list of key words that are most commonly used when cyberbullying someone, such as "idiot", "bitch", etc. (Kowalski and Whittaker 2014). However, these keywords are not entirely enough for the program to determine if someone is cyberbullying or not since these keywords can be used in any online setting. So they added modifiers to the program such as "you're", "such as", "what a _". As a result, most of the cyber aggression mostly occurred in the comment sections, roughly 80.4% occurrences.

CYBERBULLYING DETECTING METHODS: One of the current technologies to detect cyberbullying is through text-mining methods. Researchers have used machine learning models to categorize and scrape the information from famous social media. They have used natural language processing and deep learning (Bayari 2021). For NLPs, they used one of the techniques called latent semantic analysis where it is to analyze each word or phrase to see if they are positive, negative or neutral sentiment [from -1 to 1]. For example, "Pls do not think ever again"; this would have a negative sentiment score of roughly close to -1, since

this comment evokes negative emotions due to "not" combined with "ever again".

In addition, Bayari and their researchers investigated the Arabic language from a famous Youtube channel about Arabic people. They used word level and N-gram features that use SVM (support vector machines) to detect offensive comments. Then they combined all of the comments and tokenized them in order to delete all the unnecessary characters before building the layers of the model. As a result, they have found that SVMs are the most accurate classifier in classifying Arabic texts, which have an accuracy of 93%.

They have also used the Latin language to detect cyberbullying instances using a classifier called Naive Bayes. They used a balanced dataset on Twitter and fed their training data with "queer words" to classify positive comments, while "gay", "homo", and "dike" are classified as negative comments. So, the NB classifier got 67.3% accuracy (Bayari 2021).

Moreover, researchers use a common technique called bag of words in which words are collected in a "bag" of words and transformed into a vectorized word count (Perasso 2020). They do this because it is easier to use mathematics operations to perform on the textual analysis of the cyberbullying texts. When researchers use these computational tools, they found that one of the primary reasons why cyberbullying occurs is in the context of death, religion, or sexual content. This can be seen in contexts like the Israeli-Palestineans conflict where social media, as well as the news, takes advantage of the situation and that several commenters have deep emotions in either side will most likely make the commenters comment negatively.

PROBLEMS WITH CYBERBULLYING DETECTION METHODS: While these text-mining techniques and ML tools to detect cyberbullying can be quite handy, there are some technical problems associated with it. Firstly, there is the imbalance between the text, videos, and images between different social media platforms. For example, the majority of the cyberbullying text occurs via "Twitter and Whatsapp", while cyberbullying videos and images mostly occurs on Tiktok and Instagram (Perasso 2020). This imbalance might make the sentiment analysis of cyberbullying difficult because researchers would have to trace back and forth between the social media platforms and that would be time consuming. If the model predicted that the cyberbullying text occurred on Twitter, but it actually occurred on one of the videos, e.g., Instagram, that have the cyberbullying texts in them. In addition, several young individuals perceive visual cyberbullying as more dangerous compared to text, so some ML applications might not have the power capability in harnessing images and videos.

Another technical problem that might occur is due to the biases of the ML algorithms. Several comments have different context to determine whether the cyberbullying has actually occurred or not. When there is a common, shared definition of cyberbullying trained on ML algorithms, such as few relevant harmful

words, it would be difficult for the ML algorithm to detect other harmful words and will establish a new set of biases that will only select "relevant features" (Perraso 2020). Moreover, the datasets provided might not have labels or annotations on them for the data samples due to the lack of expertise done by the annotators. These can result in the risk of subjectivity biases that relies on the main definition of cyberbullying. So then, cyberbullying detection might not be correct 100% of the time because some of the comments might include crying or skull emojis or comments that may ""have been too harsh"", while in reality the 'perpetrator' was just kidding around (Heirman and Walrave 2008). An example from a Belgian school said "it was kind of a joke to me, but when I saw him at school I realized that I had driven things too far" (Heirman and Walrave 2008). Here, the model will most likely interpret the sentence as more of a negative stance, rather than a neutral or positive stance.

Moreover, even though ML algorithms can detect cyberbullying tactics, it may not detect the psychosocially of the victims of cyberbullying. In other words, it may not detect the aftermath emotions after the cyberbullying have done onto them and this might be a serious problem because the point of detecting cyberbullying is to try to stop it, as well as helping the victims. So, the impacts of these should "rely on objective and rigorous protocol of assessment" (Perraso 2020). This means that ML algorithms should not be subjective and be objective instead when assessing the severity of the cyberbullying comments.

In terms of low-funded schools, cyberbullying is even harder to detect because many students have found strategic ways "beyond the boundaries of school supervision". In addition, many victimized students stay silent, as stated earlier, and even if ML algorithms found the culprit, the likely chance is that it would be difficult to trace back why the cyberbullying occurred in the first place. The primary reason why students do not speak up is the possibility of their parents rebuking their Internet access (Heirman and Walrave 2008).

Researchers will need to update their machine learning tools to account for the psychological effects of victimized students, be able to detect cyberbullying from visual images and photos more accurately, and be able to detect cyberbullying from place to place, e.g., from Instagram to WeChat. To improve the imbalance from text to images, scholars and researchers should collaborate even further from different fields so that annotators will not have a difficult time or forget to write the important information onto their data.

Overall, cyberbullying is still an ever-growing problem that needs to be addressed as soon as possible so that no one will be harmed from this anymore. So, one of the ML techniques that needs to be developed and continue working on it is the bag-of-words algorithm (BoW), despite being the most common approach to detect cyberbullying. In addition, the psychological welfare aftereffect of the cyberbullying victims will need to be implemented across the world so that the cyberbullying detection methods will not go in vain.

ISRAEL-PALESTINE CONFLICT: The Israel-Palestine conflict, deeply rooted in religious, political, and territorial disputes, has entered a new arena in recent years — social media. In this case study, the weaponization of social media involves both sides strategically using platforms to disseminate propaganda and misinformation to influence public perception and international opinion (Loewenstein, 2023). The conflict between Israel and Palestine has been ongoing for over a century, marked by periods of intense violence, political negotiations, and shifts in control over contested areas (Loewenstein, 2023). Key historical events include the creation of the state of Israel in 1948, the wars of 1967 and 1973, the First and Second Intifadas, and numerous peace attempts and proposed two-state solutions (Loewenstein, 2023). Each side views the conflict through deeply entrenched narratives of victimhood, resistance, and survival. In the digital age, both Israeli and Palestinian actors have increasingly turned to social media to shape narratives and mobilize support. Platforms like Twitter, Facebook, Instagram, TikTok, and Telegram are used to share real-time updates, rally supporters, and spread both accurate and misleading information (Dixit, 2023).

The Israeli Defense Forces (IDF) have become adept at using social media to voice their narrative (Dixit, 2023). They employ a sophisticated and well-funded approach, leveraging these platforms to share infographics, videos, and live updates on military operations. For instance, during Operation Pillar of Defense in 2012, the IDF actively tweeted about their actions, shared videos of airstrikes, and posted infographics detailing their military successes and the threats they faced from Hamas (Loewenstein, 2023). This strategy is not just about communication but also psychological warfare, aiming to demoralize opponents and garner international support by portraying Israel as a victim of terrorism and a defender of democratic values, especially as Israel is the only democracy in the Middle East and a strategic ally for Western powers (Loewenstein, 2023). The IDF's social media efforts are coordinated, involving numerous officers and soldiers, and often resemble a well-orchestrated marketing campaign (Loewenstein, 2023). On the other hand, Palestinian groups and their supporters use social media to highlight the human cost of Israeli military actions and the conditions of life under occupation in active war zones (Loewenstein, 2023). Past campaigns have leveraged tools like hashtags, such as #GazaUnderFire, as well as images of destruction and casualties to elicit worldwide outrage (Loewenstein, 2023). However, these efforts are often less structured and rely more on grassroots mobilization and user-generated content (Loewenstein, 2023).

Both sides of the conflict, as well as external actors, use fake news and manipulated media to sway opinions and garner international support (Dixit, 2023). For example, videos and photos from unrelated events are often repurposed and misrepresented as current incidents in the conflict (Dixit, 2023). Twitter, under Elon Musk's ownership, has been particularly criticized for its handling of misinformation (Dixit, 2023). Changes to verification processes and content moderation have made it easier for false information to spread (Dixit, 2023). As such, the European Union and other entities have urged social media companies to take stronger ac-

tions against the dissemination of fake news (Dixit, 2023).

CASE STUDY: This section will highlight the ways that both sides of this conflict have leveraged social media as a weapon to gain certain advantages in the eyes of the international public, and how their strategies have aided them in shaping a public opinion that is to their benefit. Social media in this context has aided in carrying out the arrest and surveillance of Palestinians, spreading misinformation and disinformation campaigns, and helping the IDF carry out various key social media strategies (Goldstein, 2023).

The Israeli government leverages social media to monitor anti-Israel narratives on various platforms, including by cracking down on Palestinian users' social media posts (Goldstein, 2023). In fact, the Israeli government has initiated numerous prosecutions against Palestinians for their social media posts. This includes arresting students and peace activists, holding them in maximum security prisons without bail, and accusing them of inciting terrorism based on their online expressions (Goldstein, 2023). For instance, the Israeli police initiated 250 cases post-October 7 attacks, leading to around 80 indictments (Goldstein, 2023). The Israeli government has also created a specialized task force called NSO and a cyber unit for extensive monitoring and censorship of social media activity, often bypassing legal protections (Goldstein, 2023). The Cyber Unit collaborates with social media companies to remove content flagged as inciting terrorism without the required warrants and court orders (Goldstein, 2023). These efforts are often selective and disproportionately directed at Palestinians and Jewish peace activists, while similar rhetoric from right-wing extremist groups receives less attention (Goldstein, 2023). This selective enforcement has been criticized for undermining civil liberties and being highly subjective (Goldstein, 2023).

Moreover, throughout recent developments in the conflict, an array of misinformation campaigns has been spreading over Twitter, Facebook, Instagram, and TikTok. For example, videos from video games and unrelated conflicts have been misrepresented as real footage, causing mass outrage and grossly misrepresenting the state of the conflict (Dixit, 2023). A video claimed to show Hamas fighters taking down an Israeli helicopter was actually footage from a video game; Community Notes on Twitter identified the misinformation, yet the video remained up, accumulating millions of views (Dixit, 2023). Likewise, influencers on platforms such as TikTok have reported receiving bribes to support certain ideologies. "YourFavoriteGuy" on TikTok was offered $5,000 to support Israel publicly (Goldstein, 2023). This dangerous propaganda is often subject to boosting by the platforms' algorithms. According to PBS, the algorithms on social media platforms often promote extreme and disturbing content, which exacerbates the spread of misinformation (Norris, 2023). Despite efforts to increase resources to fight disinformation, these platforms have been criticized for being complicit in the spread of disinformation (Norris, 2023).

In general, the IDF, self-labeled "world's most moral army" is known to employ a variety of social media strategies to appeal to a large audience, including via the extensive use of memes (Loewenstein, 2023). During popular events such as the World Cup, the IDF used memes and other digital content thereby hijacking hashtags to spread their messages, often weaponizing Jewish trauma in the process (Loewenstein, 2023). During conflicts, the IDF and Hamas engaged in direct social media exchanges, posting alerts, updates, and taunts (Loewenstein, 2023). These efforts are supported by the involvement of at least 70 officers and 2,000 soldiers in designing, processing, and disseminating propaganda across various social media platforms (Loewenstein, 2023). Notably, during the operation "Pillar of Defense" in 2012, the IDF actively tweeted about their actions, shared videos of airstrikes, and posted infographics detailing their military successes and threats from Hamas (Loewenstein, 2023). The IDF's official Twitter account declared the start of their campaign by sharing an infographic listing Jabari's crimes "ELIMINATED" stamped across his face, and posted a YouTube video of the strike on Ahmed al-Jabari, which was viewed nearly 5 million times by November 2012 (Visser, 2012). Operation "Cast Lead" which spanned 2008-2009 also featured some of the IDF's social media strategies, including by listing the killings or arrests of Palestinian "terrorists" (Visser, 2012). These are recurring tactics which have allowed Israel to stay ahead of Palestinians in Western public view, often granting them justifications for their military offensives against the Palestinian population. Furthermore, Operation "Protective Edge" carried out in 2014 features yet another tactic, whereby the IDF featured pro-gay and pro-feminist messaging alongside militaristic content to link Israeli military actions with Western values (Loewenstein, 2023). Lastly, upon encountering negative responses to their posts and content, the IDF often floods social media with posts to drown out criticism and control the narrative.

On the Palestine end of things, Cyabra, a Tel Aviv-based social threat intelligence company, identified a highly organized influence operation featuring thousands of fake accounts (Mendez, 2024). These accounts, many inactive for years, were activated to spread disinformation about the Gaza conflict (Mendez, 2024). Another coordinated attempt by Palestinian forces was disseminating a narrative portraying Hamas members as compassionate towards hostages by clipping video segments out of context (Mendez, 2024). This content was then distributed by fake accounts to shape a favorable public perception of Hamas, which Israel, and in turn, Western media have largely labeled as terrorists (Mendez, 2024). Efforts on the part of Palestinans remain constrained to user-based single posts designed to draw attention to the lifestyle of many refugees and displaced civilians as a result of armed and violent conflicts.

Finally, the Israeli government has successfully colluded with the holding companies of various social media platforms, including Meta, TikTok, and Google (Goldstein, 2023). These companies have received thousands of requests from the Israeli government to remove content related to the conflict; having complied, a significant percentage of flagged content was removed, which critics argue amounts to state censorship laundered through private companies (Goldstein, 2023). These

companies have also showed bias in their content moderation strategies, as studies have shown that content moderation biases disproportionately affect Arabic content due to mistranslations and other factors (Goldstein, 2023). For example, Facebook repeatedly removed posts simply for using the Palestinian flag emoji during the conflict (Goldstein, 2023).

PROBLEMS WITH THE TECHNOLOGY: Social media, initially meant as a tool for users to find and join online communities that reflect their backgrounds, opinions, and topics of interest, has since evolved into a powerful tool for the large-scale spread of misinformation and manipulative agendas. In the context of the Israel-Palestine conflict, either side uses these platforms to their benefit by mobilizing support through content that tends to draw attention to specific political narratives, intentionally negating any other contradicting ideology. Whether that is via hijacking trending topics during worldwide events, like the World Cup, or denouncing the bombing of civilian areas and the taking of hostages and prisoners, this conflict is a powerful examples of the mass-scale harm that social media may bear unto the global political arena.

Attempting to address such an issue is rather challenging seeing as though current algorithm and machine learning technology have strides left to make toward accurately identifying and flagging misinformation, propaganda, bots posing as real users, and trolls who seek to stir controversy. Likewise, lobbying from the Israeli government and cyber units for the suppression of pro-Palestine content poses an immense threat to basic human rights as they apply social media, such as freedom of speech. As such, there are a range of potential solutions that should be implemented in order to mitigate damage and restore social media as a tool to connect users in a non-malicious way.

First of all, it is crucial to improve media literacy in all users, regardless of their backgrounds, age, political affiliations, etc. This includes, importantly, changes to school curricula to include media literacy-related concepts, as well as the ability to spot and fact-check suspicious claims made on the internet. Things such as fake news reporting, fact checking, and the distribution of community notes may also help on that front, but are limited without proper initial education on the topic. Second, in software development, several improvements can be made to machine learning algorithms, including the ability to read and detect visual content and improve imbalance by choosing a model that is able to better handle textual data within the scope of a broader contextual analysis. Finally, all current moderation strategies that are influenced by outside agents should be terminated immediately to ensure that user feeds are as objective and personalized as possible without any malicious contributions.

ISIS AND THE INVASION OF IRAQ: Introduction ISIS stands for the Islamic State of Iraq and Syria, and they have been an extremist terrorist organization in existence since 2013, consisting of some former Al-Qaeda elements who fought in the Iraq insurgency. Its members practice an extreme form of Islam called

Salafism, prioritizing the formation of a global caliphate that would rule over all the people under a unified structure of a modified Islamic law. This would often lead to extreme acts of violence and terror against many populations, whether it be taking over villages, cities, and towns or publishing disturbing beheading videos of journalists on the internet to shock audiences worldwide.

They reached the height of their global popularity in 2014 when they launched an invasion of Iraq, where they captured key cities Mosul and Tikrit (Thomas et al 2024). With this operation, they were able to continue capturing more areas of Iraq and Syria, where they were able to establish a caliphate over these areas for a time. Over time, ISIS was able to capture and hold control over territories in eight countries in 2015, but by 2017 they eventually lost most of it and by 2019, they lost about 95% of their captured lands as they fought United States forces and were pushed out of their regions. However, these actions that ISIS took to invade, fight, and conquer were marked by an interesting tool: social media. ISIS fighters themselves propagated the use of a new hashtag called #AllEyesonISIS, which they used to essentially announce the invasion of Iraq to the world (Brooking et al 2016). Though it may seem like an unconventional move to make, this proved to be one of the most effective tactics ISIS used to popularize themselves to the entire world and allow them to have a voice and gain traction all through weaponizing social media.

Case Study: Since 2014, ISIS has been using social media to various extents in order to achieve their goals and bring attention to themselves on a global scale. The content they create to post mainly deals with spreading their ideology through various forms of propaganda, where they are "using the Internet for psychological warfare, publicity, propaganda, fundraising, recruitment, networking, sharing information and planning" (Awan 2017). This type of content is then broadcast through various channels that are easily accessible to many people, including many youths using the internet today, whose vulnerable and impressionable minds absorb this imagery and are encouraged to join them, which has been "helping the group draw at least 30,000 foreign fighters, from some 100 countries, to the battlefields of Syria and Iraq" (Brooking et al 2016). In addition to recruitment and indoctrination endeavors, ISIS has also used social media to announce its military operations digitally, as mentioned before with their use of #AllEyesonISIS during the invasion of Iraq. This allows for a domino effect of how ISIS is creating an impact on the world, and announcing their attacks and victories was "helping to fuel a sense of the Islamic State's momentum" (Brooking et al 2016). Behind this momentum are the soldiers fighting for ISIS, and through social media, the content created attempts to humanize them, where they are seen visiting wounded soldiers in hospitals and even meeting with children on the streets to hand them various candies (Awan 2017). This, along with the rest of their content, is all used by the terrorist group to place them morally above all others and justify their extremist mission while instilling fear in others to display their strength and feeling of unstoppableness.

Regarding the actual content they create and distribute, ISIS has a strategy that fulfills the impact points described previously. All of the social media creations from official ISIS channels go through a central media office: the Al Hayat Media Centre (Awan 2017). This allows for a centralized plan and image of how the group wants to present itself and instills intimidation and fear around the globe. Such content includes many recruitment-style videos that propagate their ideology as righteous and glamorize their war efforts. These videos often have a surprising amount of effort put into them by the group to make them mimic Hollywood-style action movies where "Colors are saturated, contrasted, and crisp; subjects are kept in clear and tight focus" (Brooking et al 2016). These videos contain a wide array of elements like first-person combat videos, soldiers humanizing themselves and interacting with the civilians they supposedly protect, and even soldiers at home talking about themselves and their families.

Though these are portrayed as being completely authentic, there are many instances where these heroic and altruistic videos and pictures are actually staged, where "Immaculately staged photos, filtered through Instagram, transformed a ragtag force riding in dusty pickup trucks into something larger than life" (Brooking et al 2016). To them, it does not matter whether the images that people see are real or fake, what matters is that they are seen at all, spreading fear and intimidation on a global scale and furthering their mission. Their most popular tactic to carry this out is distributing videos of their hostages getting beheaded. These videos have gained notoriety for their gruesome nature and have solidified ISIS as a group of brutal means, but for some of their videos, the execution would be cut to black instead of showing the full event. It became clear what ISIS was doing by cutting: "The event had been filmed in such a way as to make it shareable by conventional media outlets" (Brooking et al 2016). The spread of their content is just as important as the content itself to reach the widest audience possible.

As centralized as their media production is, the way their content is spread is surprisingly decentralized, but this is the key to their success. This has an effect on how ISIS can even be covered because "Despite the willingness to limit ISIS supported media, the international community is unable to achieve its objective due to ISIS's strategy of decentralization of media" (Khawaja et al 2016). The way that their content is spread is not only through official ISIS channels on platforms like YouTube, Facebook, Twitter, and Instagram but also through many smaller accounts from fighters and supporters on these platforms and others that keep popping up. This way, ISIS always has a voice that is spreading its message and plans for all to see. These accounts are created, their views are shared, and then those accounts subsequently get banned, but nothing stops these people from simply making another account and continuing to post from there, creating a "whack-a-mole" type of situation where it is hard to track the number of ISIS-supporting social media accounts. No matter how many accounts are suspended or how quickly they are suspended, "that suspension is disruptive to terrorists but not to research or intelligence gathering" (Wright et al 2016).

In addition to having their content spread openly, their recruitment has also been made away from the public eye as "Contact with sympathizers has often been made in an open forum, and then moved to private message exchanges" (Brooking et al 2016). This further humanizes those in ISIS and those they try to recruit, as they no longer are just these people across the globe, but friends. Such communication has been done through an app that ISIS themselves developed called "The Dawn of Glad Tidings" with live updates and the latest news from the group, but was later detected and suspended (Awan 2017). No matter what ISIS does on their digital front, all of these methods have the same goals: recruit, instill fear, and place themselves above all.

The impact is felt not only with ISIS but in other groups as well. Since the success of ISIS's social media campaign has been made readily apparent, other militias have followed suit and "have created their own Facebook pages and Twitter accounts, posting openly about their targets and boasting about prior attacks" (Shea et al 2023). Their methods are similar, as they have adopted unified campaign imagery and branding, where they have even started creating public hit lists against enemies of their group. Across all platforms, groups are beginning to take advantage of the power of social media to spread their messages of hate and recruit others to join them. The same broad objectives are observed from these groups: "to overwhelm the states' adversaries" and "mobilize their own citizens and supporters and bind them to the state" (Brooking et al 2016). Social media makes this incredibly easy, as any form of content can be uploaded and observed by many people, thus allowing people to act upon these ideas and further the message of these terrorist groups.

PROBLEMS WITH THE TECHNOLOGY: As mentioned above, one of the main problems with social media and its use by ISIS is the fact that supporters of the terrorist organization can continue to spread their content no matter what methods of suspension are used against them. Many social media platforms like Twitter have made attempts at suspending many ISIS-supporting accounts, once claiming "that it had suspended 10,000 ISIS linked accounts in a single day" (Wright et al 2016). However, new accounts keep popping up and keep posting the same information that was posted before, thus allowing the group to continue their spread. Currently there are no methods of actually identifying resurgent accounts from this large amount of data, which makes it difficult to get a true grasp on how many supporters are out there. Still, knowing that at least 10,000 were suspended does give insight into the amount of ground ISIS has covered on the digital landscape. Though it would be somewhat disruptive to the terrorists themselves with suspension, it does not disrupt their messaging. New accounts are made, and the flow of information continues. By suspending their accounts, there are theoretically less terrorists using Twitter, thus preventing them from spreading their recruitment, propaganda, and threats, but there is also the concern of lack of intelligence or freedom of speech.

This is the difficulty in finding a proper solution to the issue. Finding these accounts among millions and examining their messaging can only be done as soon

as they make a post. Any account, especially when considering freedom of speech, is allowed to say almost anything, so restricting that is an ethical dilemma on its own. At least when considering deontology, the universality principle examines the universality of certain maxims that can be made, and for a maxim that allows anyone to have the freedom of speaking what is on their mind, it is hard to completely universalize when considering things like hate speech, terrorist plans, or graphic imagery. At best, it can be a caveat of sorts to freedom of speech as one cannot say certain hateful things. However, with a known terrorist organization like ISIS that has objectively committed acts of terror against innocent people and threatened to continue over time, then it becomes easy to find grounds to suspend any account supporting such an extreme cause. Suspensions can happen all they want, but creating accounts does not require too much from anyone, as they would only need to provide a name, email address, and other miscellaneous info that can easily be falsified.

5. Ethical and Societal Issues

Impact on Democracy and Political Process

The weaponization of social media, in particular regards to the dissemination of fake political news, can have a drastic effect on our nation's fundamental democratic system. As mentioned in the case study about the Russian interference in the 2016 United States presidential election, foreign actors were able to utilize the preexisting social media platforms to spread propaganda and fake news. They generated news articles and social media posts directed towards specific target demographics in order to misinform voters and sow social discord. All of this was conducted using social media and tangential technologies. These technologies raise a huge ethical concern in the way they affect voters. These platforms serve as vehicles for malicious actors to mislead voters towards decisions and policies that may not be in the best interest of themselves or their nation.

Using a deontological framework, we can highlight how this technology is being unethically to harm American citizens on the web. Duty ethics highlights the importance of adherence to moral rules when taking action. Central to this framework is the reciprocity principle. The reciprocity principle states that, "Act as to treat humanity, whether in your own person or in that of any other, in every case as an end, never as it means only"(Poel 2011). Treat other people not as a means towards a goal, but as individuals with their own values and autonomy.

Applying this principle to this case, we can see how social media can be an unethical tool that enables others to violate the ability of citizens to conduct a proper election. The actions of those who use social media to spread false information treat their tricked audience as means to promote their political goals and not as individuals with a right to undisturbed voting. The malicious attackers and technology they use violate the autonomy of the Americans as they are preventing them from making informed choices that would benefit their wellbeing. Instead,

they manipulate them into decisions that could go against their personal interest in favor of those of the unethical actors. Democracy is a system that relies on its voters to make good decisions that will help themselves and the nation grow. In falsely educating the voters, they hurt a key tenet of democracy, ultimately, leading less overall social good and hurting the individual. Additionally, when voters are aware there is an attack on their social media, this erodes trust in our democratic institution, media, and each other. This only further hurts the wellbeing of the citizens of the United States as it breaks down their belief institutions. The effects of social media can be extended to other cases like the social media warfare between Israel and Palestine. Both sides create social media campaigns with misinformation and heavy skew towards their side in order to mobilize support. Both sides violate the reciprocity principle as they treat their viewers as means to garner support for their side and ignore whether their propaganda is giving people a fair perspective on the complex situation.

The use of social media in a manner to spread fake news and attack democratic system also has meaningful social consequences. In the case of the Russian trolls of the IRA, they created accounts on both sides of online discussions and campaigns in order to create a greater social divide. Their campaigns ultimately create a polarizing environment that attacks societal harmony. It creates the illusion that the differences between people are greater than they really are. Social trust and understanding is important in high trust democratic nations where we rely on others to ensure our country's political process functions properly.

Machine Learning Algorithms and other Software for Platform moderation

As mentioned in the case study regarding cyberbullying and the Israel-Palestine conflict, one of the possible technological solutions to address bullying is the use of software tools, such as machine learning algorithms to identify undesirable content from social media platforms and remove them. The amount of cyberbullying, fake news, and other malicious content is immense. These technologies are promising in that they allow automated systems that can parse vast amounts of data to remove the unwanted media, something that would be impossible for human moderators to accomplish on large social media platforms. This, however, raises the concern of censorship. If machine learning algorithms are not full proof, it is possible that they could delete legitimate news and hinder legitimate political discussion. It is also possible that something could be widely accused of being fake news, and end up censored, when in reality it had legitimacy.

We can analyze the ethical concerns of machine learning algorithms and other software solutions through a deontological framework. Specifically, we can use the universality principle to evaluate the action of removing fake news. The universality principle is as follows: "Act only on that maxim which you can at the same time will that it should become a universal law"(Poel 2011). In other words, act according to a principle that you would want others to universally follow if they were in the

same situation. Viewed under a deontological lens, it can argued that censor algorithms are unethical as they operate under the maxim that even if a story is real, if those in control of social media platforms deems it to be fake news, they can delete its completely across social media platforms, which also serve as spaces for public discourse. Social media is a key location for discussion in the modern age, removal of a story effectively hides it from a vast portion of the population. If moderators or moderating algorithms of social media platforms mistakenly removed a new story, it not only limits the information available to the public, but it violates the first amendment right for an individual to free speech. Since social media platforms are a vital place for communication, removing their posts for illegitimate reasons can be in a sense illegally removing their voice.

The ethical consideration thus far is operating under the assumption that platforms are good actors who will not intentionally hide accurate content. If platforms can freely remove fake news stories, what are the safeguards to prevent them from purposefully suppressing true stories under the guise of protecting against fake news. Situations like these have already occurred. Near the end of the 2020 presidential election cycle, there were reports that a laptop belonging to the son of Presidential Candidate Joe Biden, Hunter Biden, was discovered and contained confidential information relating to possible illegal activity. At the time, popular sentiment and the narrative being reported by mainstream media outlets was that the story was a hoax. The story was heavily suppressed on social media platforms like Twitter and Facebook despite later being revealed to be true(Thiessen 2022). The political implications of the laptop are not in the scope of this paper, however, if this true new story was not suppressed, it is possible that voters would have shifted their opinions of certain issues or candidates. With algorithms used to monitor social media platforms used at the discretion of social media platforms, it gives them significant power to affect public discourse. This is not to say that platforms should not monitor and remove cyberbullying or the dissemination of fake news, but that there needs to be robust mechanisms that can guarantee trust in algorithms and the administrators of social media platforms. Machine learning algorithms still present a promising approach to protecting online spaces from harmful actors, but it is important we use caution in our implementation.

1. Privacy Concerns

Data Collection and Surveillance: Social media platforms collect vast amounts of personal data, often without users' explicit consent or full understanding. This data can be used for targeted advertising, sold to third parties, or potentially misused, leading to privacy invasions.

Lack of Transparency: Users are often unaware of how their data is being used, stored, or shared. The lack of transparency in data handling practices raises significant ethical concerns.

2. Mental Health Issues

Addiction: The design of social media platforms encourages addictive behaviors through features like endless scrolling, notifications, and likes, which can lead to excessive use and negative mental health impacts.

Comparison and Self-Esteem: So: Social media often presents idealized versions of life, leading to unhealthy comparisons, low self-esteem, anxiety, and depression, particularly among young users.

3. Misinformation and Fake News

Spread of False Information: Social media allows misinformation and fake news to spread rapidly, influencing public opinion and potentially causing harm. This can undermine democratic processes and contribute to societal polarization.

Algorithmic Bias: Algorithms prioritize sensational or engaging content, which often includes misleading or false information, exacerbating the spread of misinformation.

4. Polarization and Echo Chambers

Confirmation Bias: Social media algorithms often show users content that aligns with their existing beliefs, reinforcing confirmation bias and creating echo chambers. This can lead to increased polarization and intolerance of differing viewpoints.

Division and Conflict: The segmentation of online communities can heighten social divisions and conflicts, both online and offline.

Anonymity and Harassment: The anonymity provided by social media can embolden individuals to engage in cyberbullying, harassment, and hate speech, causing emotional and psychological harm to victims.

Lack of Accountability: Perpetrators of online harassment often face little to no consequences, which can perpetuate harmful behaviors.

6. Impact on Democracy and Political Processes

Manipulation and Influence: Social media can be used to manipulate public opinion and interfere with political processes. This includes the spread of propaganda, the use of bots and trolls, and foreign interference in elections.

Erosion of Trust: The spread of misinformation and the manipulation of content can erode trust in institutions, media, and democratic processes.

7. Commercial Exploitation

Exploitation of Users: Social media platforms often prioritize profit over user well-being, exploiting users' attention and personal data for commercial gain.

Advertising and Consumerism: The constant exposure to targeted advertising can

contribute to consumerism and materialistic values, impacting societal norms and personal well-being.

8. Inequality and Access

Digital Divide: While social media connects many, it also highlights and exacerbates the digital divide. Those without access to technology or the internet are left out of the digital conversation, widening societal inequalities.

Unequal Representation: Marginalized communities may be underrepresented or misrepresented on social media platforms, perpetuating stereotypes and inequalities.

Ethical Considerations

Consent and Autonomy: Ethical concerns arise regarding the extent to which users can provide informed consent for data collection and understand the implications of their social media use. Responsibility and Regulation: The ethical responsibility of social media companies to protect users and ensure the accuracy of information is a contentious issue. The need for regulation versus the importance of free speech is an ongoing debate. Addressing these issues requires a multifaceted approach, including better regulation, improved transparency, user education, and the development of ethical guidelines for technology use.

6. Recommended Solutions

Improve media literacy

As outlined in the cases on the 2016 Presidential Election, the Israeli and Palestinian war, and the Isis propaganda, social media is an effective tool for bad actors to disseminate fake news and false narratives. Although there are possible technological solutions to address these social media posts, such as machine learning models, much of the technology is still developing and there is still time before full integration into social media platforms. Instead of reinforcing the robustness of our social media systems against fake information campaigns, another solution is to build up the necessary defense in social media users to help them recognize and accurately parse malicious social media content.

Improving media literacy in the population is another approach with provable results that can help attack fake news campaigns. According to the Institute of Museum and Library Services, media literacy can be defined as “the skills associated with using technology to enable users to find, evaluate, organize, create, and communicate information; and developing digital citizenship and the responsible use of technology” (IMLS 2010). Media literacy skills are crucial for individuals to engage with the digital world properly. Increased education in digital media literacy can help individuals better recognize fake news headlines. A study conducted by

Andrew M. Guess, an assistant professor in politics and public affairs at Princeton University highlighted the importance of digital media literacy for people engaging with online news articles and political posts. Their experiment showed that after social media intervention, participants in the United States showed a 26.5% improvement in ability to discern between mainstream and false news headlines (Guess 2020). Other studies showed similar results supporting this claim. A paper from the MIT Sloan School of Management found that regardless of age, race, gender, education, or political partisanship of participants, more digital media literacy education led to an increased ability of participants to discern between real and fake news posts. Participants showed an increased ability to recognize fake news, whether it be political content or news related to Covid 19(Sirlin 2021).

Curricula for effective digital media literacy focuses on the development of several skills such as critical thinking, communication, analyzing information, self-awareness, and problem-solving, and navigating systems (LINCS 2019). Successful digital literacy curriculum should instill in students the necessary analytical skills to think critically about the content they engage with online. There has been a larger push in recent years for digital media literacy education to be introduced in schools, especially as today's youth are exposed to digital technologies at an early age. Many schools have already begun implementing these programs into their curriculum. This shifting of curriculum is not only in America, but also all around the world. Estonia, a former part of the Soviet Union with a sizable minority of Russian speakers, regularly receives propaganda from Russian news and trolls. In order to address these issues, they have introduced more digital media education in addition to their already advanced technology curriculum (Robbins 2020).

Increased digital media literacy is an ethical solution to the issue of weaponization of social media in particular regards to fake news. The ethical framework of utilitarianism argues that the most ethical action is the one that maximizes the overall good and minimizes suffering. Following this principle, digital media literacy is an ethical solution because it maximizes happiness by enabling individuals to make more informed decisions and gain trust in media. When citizens develop a better understanding of digital media, they can make more informed decisions regarding their political or health choices. The collective benefit of a more knowledgeable and awareness leads to better decision making discourse, maximizing social benefit. In addition, a better understanding of news enables greater trust in news media as individuals have applied their knowledge to verify the post or news. This trust is necessary for a democracy and a restoration of this trust ensures citizens make informed decisions, leading to better society.

Improve ML and algorithm technology

Now, we have seen throughout this research paper how machine learning techniques and algorithms have tried to combat the weaponization in social media through SVMs performing semantic analysis, using a balanced dataset to classify hate comments, and bag of words where each comments go through a mathematical

vector so that the model can learn and predict what are the future comments going to say. There are a variety of improvements to improve ML algorithms so as to increase the accuracy of detecting bots, trolls, and cyberbullying in social media.

The Usage of TensorFlow: TensorFlow is a ML algorithm that "employs deep learning networks" that are designed to do better in pattern recognition through "training sensory data" (Cartwright, et. al 2019). It is then inputted into a neural network while its weights are being calculated. In other words, the comments, tweets, and replies in social media will be put into a neural network, calculate its weights, and output a vector that will be analyzed by a team of researchers. This output vector will then be used to analyze a set of words using SentiStrength, an algorithm that assigns "positive or negative" values to the "lexical units in the [comments]" (Cartwright, et. al 2019). In addition, SentiStrength also adds in the emotional or attitude value in each of the texts and outputs them into each of their own categories, such as positive, negative, or neutral.

Random Forest Classification: RFCs are a model that average a significant, large amount of dataset samples to improve the model's predictive accuracy and reduce overfitting. For context, overfitting is when a model learns the data too well and does not generalize the data well enough, meaning it will most likely predict what comments are going to say most of the time and not look at the surrounding context that may be meaningful.

By using these recommendations to improve machine learning techniques, it will ensure that researchers and ML practitioners can do the best job in analyzing sentiment analysis of each comment, tweets, and text conversations.

7. Conclusion

Ultimately, this report examines the weaponization of social media to shape public opinion, focusing on the use of trolls, bots, and other malicious agents to influence users, particularly in the realm of their political opinions and beliefs. Through four case studies—cyberbullying, the Israel-Palestine conflict, ISIS in Iraq, and the 2016 US Presidential Election—it provides a socioethical analysis of social media's role in disseminating sensitive and controversial content. Social media platforms have revolutionized global communication, but their misuse for political manipulation and the spread of misinformation has become a significant concern. Trolls and bots spread false information, disrupt political processes, and amplify controversial content, as seen in the 2016 US Presidential Election. These issues highlight the need for measures to mitigate the impact of misinformation campaigns. Originally intended for personal interaction, social media platforms have been exploited for political mobilization and influencing public opinion, with examples like the Arab Spring and the 2016 US Presidential Election underscoring their dual nature and the necessity for comprehensive regulation and intervention. There are several social and ethical consequences of social media and related technologies. The weaponization of social media to spread fake news threatens our democracy and violates the rights of our citizens. The use of monitoring algorithms on our

platforms raise ethical questions about censorship and freedom of speech. The solutions we propose include different machine learning models and increased digital media literacy education. As we continue to digitize and advance in fields like artificial intelligence, the integration of the internet and social media into our daily lives will only deepen. The increasing use of social media will inevitably attract more scrutiny. The recent conflict between TikTok and the United States Congress highlights that the social media platforms and related technologies will undergo rigorous evaluation, leading to efforts to enhance these platforms and assess their impacts.

8. References

Awan, Imran. "Cyber-extremism: Isis and the power of Social Media." Society, vol. 54, no. 2, 15 Mar. 2017, pp. 138–149, [https://doi.org/10.1007/s12115-017-0114-0]

Barrett D, Horowitz S, February 18, 2018, Russians indicted in 2016 election interference, Washington Post, [https://www.proquest.com/news/docview/2002735461/FC276EF46224173PQ/15?sourcetype=Newspapers]

Barsotti, S. Heinz College, Carnegie Mellon University. October 2018. Troll farms and fake news: The weaponization of social media. [https://www.heinz.cmu.edu/media/2018/October/troll-farms-and-fake-news-social-media-weaponization].

Bayari Reem, & Bensefia, Ameur. (2021). Text Mining Techniques for Cyberbullying Detection: State of the Art. Astes, 6(1). [http://dx.doi.org/10.25046/aj06018] Brooking, Emerson, and P.W. Singer. "War Goes Viral: How Social Media Is Being Weaponized across the World." The Atlantic, Nov. 2016. [https://www.theatlantic.com/magazine/archive/2016/11/war-goes-viral/501125/]

Cartwright, B., Weir, G. R., Frank, R., & Padda, K. (2019). Deploying artificial intelligence to combat disinformation warfare. public opinion, 22(24), 25. [https://www.researchgate.net/profile/Karmvir-Padda/publication/364372276_Deploying_Artificial_Intelligence_to_Combat_Disinformation_Warfare_Identifying_and_Interdicting_Disinformation_Attacks_Against_Cloud-based_Social_Media_Platforms/links/63501af66e0d367d91aba7a2/Deploying-Artificial-Intelligence-to-Combat-Disinformation-Warfare-Identifying-and-Interdicting-Disinformation-Attacks-Against-Cloud-based-Social-Media-Platforms.pdf]

Congressional Digest. Russian Interference in the 2016 Election: Mueller Report Oct2019, Vol. 98 Issue 8, p9-10. 2p. [https://web.p.ebscohost.com/ehost/detail/detail?vid=4&sid=8bb7c32b-3f5e-47f9-87ef-3c90e4df8efa%40redis&bdata=JnNpdGU9ZWhvc3QtbGl2ZQ%3d%3d#AN=138854420&db=f5h]

Davies, D. NPR. October 9, 2018. The weaponization of social media and its real-world consequences. [https://www.npr.org/2018/10/09/655824435/the-weaponization-of-social-media-and-its-real-world-consequences].

De Angelis, J., & Perasso, G. (2020). Cyberbullying detection through machine learning: Can technology helps to prevent internet bullying?. International Journal of Management and Humanities, 4(57), 10-35940. [https://www.ijmh.org/wp-content/uploads/papers/v4i11/K10560741120.pdf]

Digital Literacy, 2019 Literary Information Communication System [https://lincs.ed.gov/state-resources/federal-initiatives/teaching-skills-matter-adult-education/digital-literacy]

DiResta R, October 10, 2019, The Tactics & Tropes of the Internet Research Agency, New Knowledge, [https://digitalcommons.unl.edu/senatedocs/2/]

Dixit, P. Al Jazeera. October 10, 2023. Social media platforms swamped with fake news on the Israel-Hamas war. [https://www.aljazeera.com/news/2023/10/10/social-media-platforms-swamped-with-fake-news-on-the-israel-hamas-war].

Durso, T. Princeton University. July 28, 2023. Social media polarization and the 2020 election: Insights from SPIA's Andrew Guess. [https://www.princeton.edu/news/2023/07/28/social-media-polarization-and-2020-election-insights-spias-andrew-guess-and].

Goldstein, L. The American Prospect. December 12, 2023. Palestinians imprisoned for social media posts. [https://prospect.org/world/2023-12-12-palestinians-imprisoned-social-media-posts/].

Guess A. M, April 28, 2020, A digital media literacy intervention increases discernment between mainstream and false news in the United States and India, Proceedings of the National Academy of Sciences of the United States of America [https://www.pnas.org/doi/epdf/10.1073/pnas.1920498117]

Harris, S, October 31, 2017, The Mueller Investigation: Mueller Case Makes New Link to Russia, Wall Street Journal, Eastern edition [https://www.proquest.com/news/docview/1957653745/FC43CDF8AEAD4E78PQ/4?parentSessionId=n0x1LwAOn9j3ldpVdY7QDf5QxILz80SlLS3wnemTdhY%3D&sourcetype=Newspapers&parentSessionId=9pPhqFd%2FCSeoyOLtXY2Iy4BZS10cz2fQY8PjG3C1rBA%3D]

Heirman, W., & Walrave, M. (2008). Assessing concerns and issues about the mediation of technology in cyberbullying. Cyberpsychology, 2(2). [https://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtyp e=crawler&jrnl=18027962&AN=43315196&h=mrQbiUP%2FsVQcAtratOGVjQTHZ7y bOwl76xpjeMKNNUzu03Qdc%2BhWh%2FnzhYWv7aK-

WzjkKM1raJnorSyI5%2Bd5D EQ%3D%3D&crl=c]

Khawaja, Asma Shakir, and Asma Hussain Khan. "Media Strategy of ISIS: An Analysis." Strategic Studies, vol. 36, no. 2, 2016, pp. 104–21. JSTOR, [https://www.jstor.org/stable/48535950]

Krever M, Chernova A, February 14, 2013, Wagner chief admits to founding Russian troll farm sanctioned for meddling in US elections,CNN [https://www.cnn.com/2023/02/14/europe/russia-yevgeny-prigozhin-internet-research-agency-intl/index.html]

Loewenstein, A. Verso Books. May 14, 2021. Social media is a warzone: The IDF's strategy for war as online spectacle. [https://www.versobooks.com/blogs/news/social-media-is-a-warzone-the-idf-s-strategy-for-war-as-online-spectacle].

Mendez, N. Wired. May 24, 2021. Gaza's social media war: Palestine and Israel's digital conflict. [https://wired.me/business/social-media/gaza-social-media-war-palestine/].

Norris, C. PBS NewsHour. October 10, 2023. Social media companies criticized as Israel-Hamas war misinformation spreads rampantly. [https://www.pbs.org/newshour/show/social-media-companies-criticized-as-israel-hamas-war-misinformation-spreads-rampantly].

Oxford University. January 13, 2021. Social media manipulation by political actors is an industrial scale problem: Oxford report. [https://www.ox.ac.uk/news/2021-01-13-social-media-manipulation-political-actors-industrial-scale-problem-oxford-report].

Poel, I, Royakkers, L., 2011, Ethics, Technology, and Engineering, Wiley-Blackwell. [https://cdn.prexams.com/6229/BOOK.pdf]

Public Law 111th Congress, December 22, 2010, Museum and Library Services ACT, [https://www.congress.gov/111/plaws/publ340/PLAW-111publ340.pdf]

Pyżzalski, Jacek. (2014). From cyberbullying to electronic aggression: typology of the phenomenon. From cyberbullying to electronic aggression: typology of the phenomenon, [https://www.taylorfrancis.com/chapters/edit/10.4324/9781315656694-7/cyberbullying-el ectronic-aggression-typology-phenomenon-jacek-py%C5%BCzalski]

Robbins J, September 23, 2020, Countering Russian Disinformation, Center for Strategic and International Studies, [https://www.csis.org/blogs/post-soviet-post/countering-russian-disinformation]

Saletta M, Stearne R, July 11, 2021, Understanding Mass Influence, Defence Science Partnering Multi-Party Collaborative Project Agreement, [https://www.unsw.edu.au/content/dam/pdfs/unsw-adobe-websites/canberra/research/defence-research-institute/2023-02-Understanding-Mass-Influence---A-case-study-of-the-Internet-Research-Agency.pdf]

Shea, Joey, and Ruba al-Hassani. "Hate Speech, Social Media and Political Violence in Iraq: Virtual Civil Society and Upheaval." The Tahrir Institute for Middle East Policy -, 22 Feb. 2023, [timep.org/2021/02/11/hate-speech-social-media-and-political-violence-in-iraq-virtual-civ l-society-and-upheaval/]

Sirlin N, December 6, 2021, Digital literacy is associated with more discerning accuracy judgments but not sharing intentions, Harvard Kennedy School Misinformation Review, [https://misinforeview.hks.harvard.edu/article/digital-literacy-is-associated-with-more-discerning-accuracy-judgments-but-not-sharing-intentions/]

Stricot, M. CNRS News. October 31, 2017. How social networks manipulate public opinion. [https://news.cnrs.fr/articles/how-social-networks-manipulate-public-opinion].

Thiessen M. A, December 9, 2022, The Suppression of Hunter Biden's laptop is a huge scandal, Washington Post, [https://www.washingtonpost.com/opinions/2022/12/09/hunter-biden-laptop-suppression-twitter-fbi-social-media]

Thomas, Clayton, and Abigail Martin. "The Islamic State: Background, Current Status, and U.S. Policy." Congressional Research Service, 6 May 2024, [crsreports.congress.gov/product/pdf/IF/IF10328]

Timberg C, November 25, 2016, Russian propaganda effort helped spread 'fake news' during election, experts say: Researchers say sophisticated tools were used to boost Trump and undermine Clinton, Washington Post, [https://www.proquest.com/news/docview/1843020400/84571090EAF2408APQ/1?%20Websites&sourcetype=Blogs,%20Podcasts,%20]

Visser, J. National Post. November 14, 2012. Ahmed Jabari eliminated: Israel goes to war with Hamas on social media, live tweets assassination, posts proof on YouTube. [https://nationalpost.com/news/world/israel-middle-east/ahmed-jabari-eliminated-israel-goes-to-war-with-hamas-on-social-media-live-tweets-assassination-posts-proof-on-youtube].

Whittaker, E., & Kowalski, R. M. (2014). Cyberbullying Via Social Media. Journal of School Violence, 14(1), 11–29. [https://doi.org/10.1080/15388220.2014.949377]

Wright, Shaun & Denney, David & Pinkerton, Alasdair & Jansen, Vincent & Bryden, John. (2016). Resurgent Insurgents: Quantitative Research Into Jihadists Who Get Suspended but Return on Twitter. Journal of Terrorism Research. 7. 1. 10.15664/jtr.1213. [https://www.researchgate.net/publication/303324774_Resurgent_Insurgents_Quantitati e_Research_Into_Jihadists_Who_Get_Suspended_but_Return_on_Twitter]

Department of Engineering, University of California, Los Angeles, Westwood, CA 90024
Email address: clyde0513@g.ucla.edu