Modes of Tracking Mal-Info in Social Media with AI/ML Tools to Help Mitigate Harmful GenAI for Improved Societal Well Being

Andy Skumanich[1]*, **Han Kyul Kim**[2]*
[1]Innov8ai
[2]University of Southern California
askuman@innov8ai.com, hankyulk@usc.edu

Abstract

A rapidly developing threat to societal well-being is from misinformation widely spread on social media. Even more concerning is ”mal-info” (malicious) which is amplified on certain social networks. Now there is an additional dimension to that threat, which is the use of Generative AI to deliberately augment the mis-info and mal-info. This paper highlights some of the ”fringe” social media channels which have a high level of mal-info as characterized by our AI/ML algorithms. We discuss various channels and focus on one in particular, ”GAB”, as representative of the potential negative impacts. We outline some of the current mal-info as an example. We capture elements, and observe the trends in time. We provide a set of AI/ML modes which can characterize the mal-info and allow for capture, tracking, and potentially for responding or for mitigation. We highlight the concern about malicious agents using GenAI for deliberate mal-info messaging specifically to disrupt societal well being. We suggest the characterizations presented as a methodology for initiating a more deliberate and quantitative approach to address these harmful aspects of social media which would adversely impact societal well being.

The article highlights the potential for ”mal-info,” including disinfo, cyberbullying, and hate speech, to disrupt segments of society. The amplification of mal-info can result in serious real-world consequences such as mass shootings. Despite attempts to introduce moderation on major platforms like Facebook and to some extent on X/Twitter, there are now growing social networks such as Gab, Gettr, and Bitchute that offer completely unmoderated spaces. This paper presents an introduction to these platforms and the initial results of a semiquantitative analysis of Gab’s posts. The paper examines several characterization modes using text analysis. The paper emphasizes the developing dangerous use of generative AI algorithms by Gab and other fringe platforms, highlighting the risks to societal well being. This article aims to lay the foundation for capturing, monitoring, and mitigating these risks.

Introduction

There is a rapidly growing threat to societal well being coming from the active propagation of *malicious-info* from various social media channels. An increasingly deleterious aspect is from the use of GenAI to amplify the mal-info. It is well known that social networks have had a profound impact on the way we communicate, share information, and interact with each other. They have become a central part of modern society, enabling people to connect with each other regardless of their location, share their thoughts and opinions, and participate in online communities (Osman, Barbaro, and Skumanich 2023). However, the rise of social networks has also led to a number of challenges, including the deliberate spread of disinformation (Barbaro and Skumanich 2023), cyberbullying (Giumetti and Kowalski 2022), and the propagation of hate speech and extremist ideologies (Govers et al. 2023), along with in general what we call ”mal-info”. In particular the latter, is being deployed for *deliberate disruption of society with the intent to cause conflict and discord*. It is essential to develop strategies to observe, monitor, and mitigate this *harmful impact* of social media. Although there have been some attempts by governments (e.g. the EU), to introduce some form of moderation with the help of the better-known platforms such as Facebook or X/Twitter, these have been limited in scope. Furthermore, various groups have turned to alternative social networks that offer a different set of guidelines and principles, such as Gab, Gettr, and Bitchute. These channels have gained popularity among certain segments of the population because they freely allow any manner of speech. While these networks offer a space for so-called *free expression*, they also raise questions about the impact of social networks in shaping perceptions and attitudes, and the potential consequences of this influence on societal well-being. Despite these many concerns, very little research has focused on these new social networks (Peucker and Fisher 2023), instead mostly concentrating on X/Twitter (Govers et al. 2023). This paper presents a preliminary step to setting a foundation for monitoring the mal-info which can cause social degradation. In addition, there is an impending issue that these actors can implement AI and GenAI for furthering the mal-info elements to degrade social wellbeing. The main findings are: (1) we present modes of characterizing the output of these channels for capturing and tracking mal-info; (2) we highlight developing implementations of GenAI by these domains. We introduce their notion of aggressive societal degradation by *Accelerationism* and how this can is exacerbated by social media. We flag the dangers of GenAI to this end.

In this article, we present characterization of three of these

[*] These authors contributed equally.

new (fringe) social networks, namely Gettr, Bitchute, and Gab. Then we conduct a qualitative analysis of posts from Gab containing a representative inflammatory term. Next and importantly, we present the new developments driven by these platforms for using generative algorithms (GenAI). Finally, we discuss the risks for the society of such algorithms, specifically to instigate social disruption and discord. We conclude by alerting the AI community to the need to develop modes of Capture, Track, Respond to address these risks in order to preserve societal well being.

New Social Networks With Reduced Moderation

New social networks were created in response to perceived censorship and moderation on established social media platforms. Some users feel that their ”freedom of speech” is being limited on platforms X/Twitter and Facebook and that their content is being ”unfairly” targeted or removed. This is a broader discussion beyond the scope of this paper as it is part of the determination of what constitutes *free speech*. As shown by (Stocking et al. 2022), more and more Americans are using these platforms for news (6% in 2022). Using tools like Similarweb[1], this number will have doubled in 2023, and is likely to continue to grow.

These new platforms have little to no moderation and lenient content policies. Although this allows for a wider range of opinions and viewpoints to be shared, however it provides a platform for mal-info. These channels can be legitimately criticized for allowing hate speech, harassment (Abarna et al. 2022), and misinformation to spread unchecked. In this domain of new no-restrictions social networks, three seem to emerge as frequently used channels, namely Gab, Gettr and Bitchute.

Gab is a social networking platform launched in 2016 and bills itself as an unfettered speech (so-called *free speech*) alternative to mainstream social media sites. It was created in response to the perceived censorship of conservative views on traditional social media sites. Gab allows users to post messages called ”gabs,” share photos, and interact with other users. It has been observed as being a platform for hate speech and far-right extremism.

Gettr is a newer social media platform that was launched in 2021 by former President Donald Trump’s senior adviser, Jason Miller. It is marketed as a ”cancel-free” platform that supports unfettered speech and allows users to share their opinions without any type of moderation. Gettr’s features are similar to those of Twitter, allowing users to post short messages called ”gettrs,” share photos and videos, and interact with other users.

Bitchute is a video-sharing platform that was launched in 2017. It was created in response to perceived censorship of fringe or provoking views on traditional video-sharing sites like YouTube. Bitchute allows users to upload, share, and view videos on various topics, including news, politics, and entertainment. It has been observed as being a platform for conspiracy theories and hate speech.

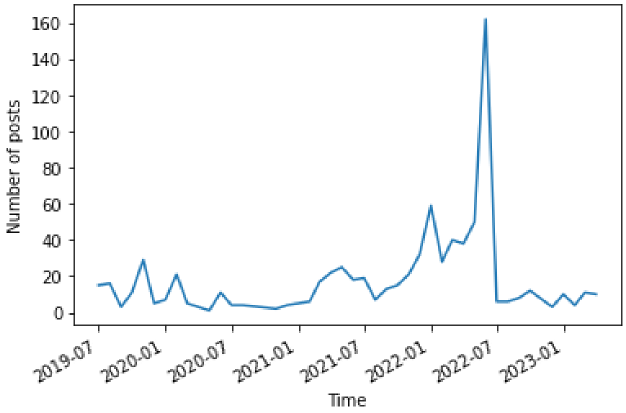

Figure 1: Timeline of posts containing $\#Cuckservative$

The main difference between these platforms is their focus and features. Gab and Gettr are primarily social media platforms that allow users to share short messages and interact with other users. Bitchute, on the other hand, is a video-sharing platform that allows users to upload and view longer-form content. Additionally, Gab has been associated with far-right extremism, while Gettr is marketed as a nominally more mainstream platform. All three platforms have been observed to tolerate hate speech and conspiracy theories, and these expressions can be used to attack the societal well-being when invoking a call to action.

Analysis of Gab

To showcase that this fringe social network does contribute to the spread of hate speech, we selected a term from the Glossary of Extremism from The Anti-Defamation League (ADL)[2]. ADL is an international non-governmental organization based in the United States specializing in civil rights law (historically focusing on anti-semitism). In their glossary, they provide an overview of many of the terms most frequently used by a variety of extremist groups and movements.

Based on this lexicon, we selected the term *Cuckservative*. In 2015, alt-right white supremacists began disparaging members of the conservative movement with the derogatory term *cuckservative*, a combination of *conservative* and *cuckold*, to describe a white conservative who putatively promotes the interests of Jews and non-whites over those of whites. The Groypers, a group that attracts white supremacists and other far-right activists, also employ the term.

For the purposes of this paper, as no API is available for Gab, a tailor-made scraper based on Python’s Selenium library (Salunke 2014) was used to automate and scale up this process. Using the $\#Cuckservative$, we retrieved 788 messages from July 4, 2019, to April 22, 2023. Figure 1 shows the number of posts over time. We can observe a peak around May-June 2022 and we could link this peak to the Mass Shooting in Uvalde, Texas and the reaction it created

[1] https://www.similarweb.com/

[2] https://extremismterms.adl.org/

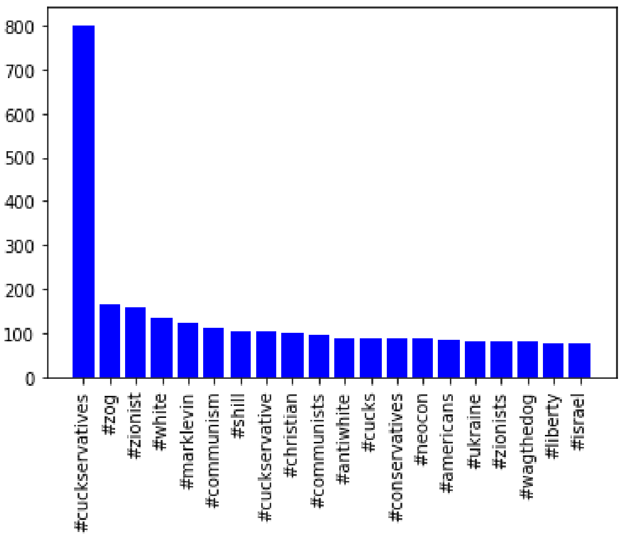

Figure 2: Top # in post

in the political sphere and for white supremacists and other far-right activists. At the political level, more than 600 conservatives[3], mostly in Texas, called for gun reform. Instead, the segment of white supremacists and other far-right activists used the term *Cuckservative* to describe these politics of attempting to develop a gun control solution and to defend their rights to carry a weapon with unfettered access.

Figure 2 shows the fifteen most used hashtags. It is interesting to note that the hashtags used for the term *Cuckservative* are in agreement with the definition by ADL. Users post using mostly hashtags concerning the *Jewish population* and their feeling about *white racism*.

In order to analyze posts, we applied a classical text pre-processing (Barbaro 2022). We removed all non-alphabetical characters (numbers, punctuations, . . .) and stopwords. Then we applied lemmatization. Afterwards, we selected texts from March to August 2022 to understand the different posts around the peak of May-June 2022.

Figure 4 shows the most used words by month. As expected, the posts make extensive use of the terms *cuckservative* and *white* over time in support of their ideology. It is interesting to note that the posts take up current issues. In March 2022, one of the main topics is the war in Ukraine. For the months of May-June, we can observe that the posts are directed against the American right, notably with the word *rat*, because according to them they betray their ideology. Finally, it is interesting to observe a certain consistency in their anti-Semitic language, with the words *jew* and *israel*. Figure 3 shows a post published on June 12, 2022, highlighting all these topics.

As a contrast, we show a similar analysis of Twitter/X for similar terms (Figure 5). There is a significant difference in the signatures of the two, showing that this approach can be used to discern useful information for monitoring. A key observation is that the levels of mal-info can be substantial, and the growth or decrease can be monitored.

#PatriotFront Proves One Thing: #CUCKS are RAT SNTICHES FOR THE #ZOG!#Conservatives will side with the #FBI because #CNN says you are a racist!Why? #CUCKSERVATIVES are slaves to JEWS NIGGERS AND FAGS!Now little quiz – Who ’forces’ us to ’tolerate’ this INSANITY:1.) The jewish Media2.) The Jewish controlled parties, bureaucracy and courts3.) The Highly Armed Judaized Police4.) the Cult of Noahides that believe rat faced murderers of God are ’G-ds’ in factor5.) All of the AboveKeep the Faith Brother ... These people need to be CRUSHED ... OUR OWN SIDE NEEDS MADE LOYAL AND TRUE before we have any enemies outside the RIGHT!

Figure 3: Example of Post published the 12 June 2022

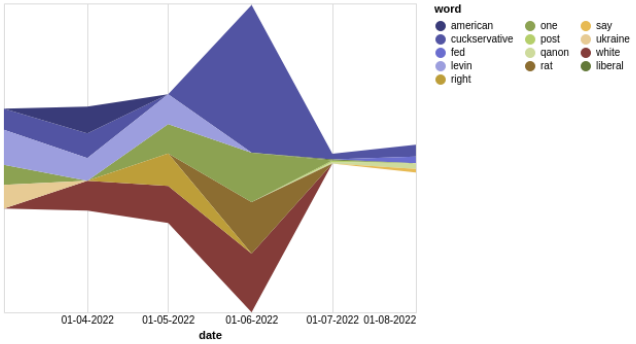

Figure 4: Streamgraph of most used words in posts by month: Gab

In order to do a deep dive into the different narratives spread by these posts, we applied the method of *keyness* analysis (Gabrielatos 2018). This approach from the fields of corpus linguistics and corpus-based discourse analysis is directed at identifying *key* items (e.g. words) in a target corpus in relation to a reference corpus based on the frequencies of items in both corpora. As such, a *keyness* analysis can support an exploratory approach to texts that gives an indication of their *aboutness*. The keyness metric chosen for this paper is that of Log Ratio, which is defined as the *binary log of the ratio of relative frequencies* (Hardie 2014). This gives a measure of the actual observed difference between two corpora for a key item (rather than a measure of statistical significance). The advantage of this is that it allows for the sorting of items by the size of the actual frequency difference between the corpora, enabling us to find the top N most key items.

Table 1 shows the key items by month. It is useful to note the different changes in narratives. For example, in March, users are mostly talking about *Covid* by using words such as *wuhan* and *thecurrentthing*. Also, they are referring to the war in Ukraine with the word *quagmire*. Then, in May-June, users seem to react to the Mass shooting and the po-

[3] https://www.reuters.com/world/us/more-than-600-conservatives-mostly-texas-call-gun-reform-2022-06-08/

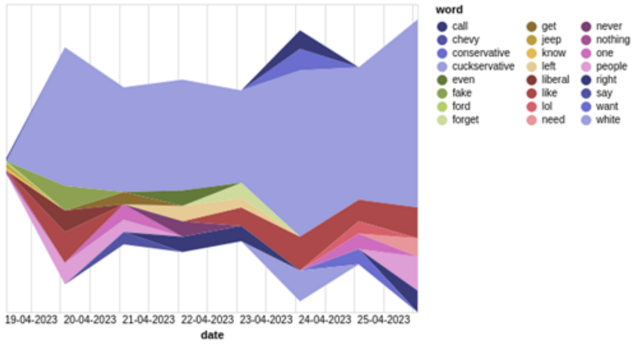

Figure 5: Streamgraph of most used words in posts by month: Twitter

Month	Key items
March	wuhan, thecurrentthing, quagmire
April	regimechange, extortion, falseflags
May	withheld, outlawed, unum
June	snitch, plunder, accountability
July	shitlibs, thinktank, fedbois
August	qanoncuck, adl, aipac

Table 1: Key items by month from March 2022 to August 2022

litical reaction in favour of gun control using terms such as *withheld* or *unum* and criticizing people betraying their ideology (*snitch*). Finally, in July-August, they return to more usual narratives (in their ideology) by criticizing The Establishment (*shitlibs*), anti-Semitic organizations such as *adl* or *aipac*[4] and other far-wing ideology like *Qanon*.

Generative AI Concerns in the Context of Threats to Societal Cohesion

In the previous section, we saw how an unmoderated social network can propagate messaging hatred against a targeted population or conspiracy ideas. We developed some modes using AI/ML for extracting signatures which provide a means to characterize the mal-info and track it in time. In this section we highlight the growing threat of the use of AI for engendering mal-info. We emphasize that this requires the monitoring tools indicated in the prior sections.

In addition to the dangers of unmoderated content, the (mis)use of AI systems in social networks also presents a significant risk to societal well being. The cultural biases present in widely used AI systems can degrade the quality of social engagements and lead to conflict. As humans increasingly rely on AI-generated content to make decisions, AI systems will have an enormous amount of influence to shape human perceptions and manipulate human behavior. To establish the extent of the dangers we highlight the views of the Gab CEO, Andrew Torba. He has explicitly stated his wishes to propagate his ideologies through his platform. He published an article on 1/27/2023 entitled: ”Christians Must Enter the AI Arms Race”[5]. In his article, Torba discusses the potential for building a new AI system that is not ”skewed” with a *liberal/globalist/talmudic/satanic* worldview like many current AI systems. He argues that if the enemy is going to use AI for evil, then they should build an AI system for good. He suggests that if people with his same ideology don’t build their own AI system, then their enemies will dominate this space and use it as a weapon against the minds of the people. This *influencer* (by way of his expanded social media presence) believes that they need to develop their AI system for the glory of God that can communicate the Truth of the Gospel to millions of people. This presents an Orwellian inversion of the notions of good and evil, and highlights the very nature of the potential for harm to societal well-being. As further evidence of the potential for harm, the Gab platform has enabled extremist and other problematic content, including verified accounts that share posts praising White Supremacy, posts espousing QAnon conspiracy theories, and hateful posts towards marginalized groups. ADL’s Center on Extremism (COE) has found multiple examples of extremist and harmful content. Torba’s overall messages support ***accelerationism***, [6] a term used by white supremacists that expresses their desire to intensify societal conflicts and collapse. This is the context where the use of GenAI is being exploited to further the elements which will lead to societal discord. The core of accelerationism is the goal of creating societal chaos. As a representative example: ”True change ... only arises in the great crucible of crisis. A gradual change is never going to achieve victory. Stability and comfort are the enemies of revolutionary change.” Accelerationism is a popular topic in private chat rooms frequented by white supremacist groups like Atomwaffen and The Base. These virtual spaces are full of discussions about steps to take to hasten the ultimate collapse. Atomwaffen articulates a white supremacist ideology rooted in nihilism and accelerationist beliefs. Violence, chaos, and destruction are themes echoed throughout their posts, propaganda, and messaging. These virulently mal-info sources are the very essence of threats to societal well-being and social cohesion.

The Gab development of a soon-to-be-available Text Generation AI called Based AI[7] is part of a comprehensive plan to develop tools for accelerationism-minded people. Gab already launched a service for Image Generation called Gabby[8] and a service for Movie Generation called Mel[9]. These features show explicitly that the intentions to use GenAI are indeed in play. The social media gives this an outsized presence.

[4] https://www.aipac.org/

[5] https://news.gab.com/2023/01/christians-must-enter-the-ai-arms-race/

[6] https://www.adl.org/resources/blog/white-supremacists-embrace-accelerationism

[7] https://gab.com/basedai

[8] https://gab.com/AI

[9] https://gab.com/movie

Discussion and the Need for Capture/Track/Respond (CTR) Modes

In this paper we provide some characterization modes of Social Media which can be expanded on to Capture, Track, and (potentially) Respond (CTR) to mal-info. The modes are initial and indicative and we intend to further develop the elements. We instead use this introduction paper to emphasize the dangers of the GenAI elements of these channels. The key point is that with the developing dimension of GenAI driven mal-info it will be essential to have methods for CTR. *Although it may not be possible to prevent the GenAI based mal-info, at least it can have CTR quantitative and qualitative monitoring and allowing for possible mitigation*. It is already apparent that online harassment can lead to actual physical harassment. In this context the potential for amplification of mal-info by the use of GenAI poses serious issues. We emphasize the importance of studying the new social networks that were fringe but are becoming increasingly important as having an out-sized impact on diminishing societal well-being as a deliberate focus of their activities. We have shown using an example of analysis of Gab's posts, signatures of how this unmoderated network propagates deliberate mal-info. We have observed that in these fringe channels there is less dilution of the mal-info so that the percentage level of mis-info, dis-info, and mal-info is observably higher vs e.g. Twitter/X. Our key findings are that (1) we highlight the growth of social media sites which promulgate mal-info; (2) we present modes of developing signatures of these channels which can allow for monitoring; (3) we emphasize the concern for societal well-being given their drive to employ GenAI to augment the mal-info messaging, with a particular concern about accelerationism; (4) we provide a basis for further research on AI modes to mitigate the issues. By having characterization modes, then the impacts of *countering* the messaging can also potentially be assessed. Some counter-messaging may be more impactful and may show up in the CTR analysis, as e.g. a drop in signatures of repeated mal-info after a targeted positive messaging campaign. This last point is the crux, which is that societal well-being requires a constantly well informed populace.

The key implications are that there should be an active Capture, Tracking, and Responding, (CTR), for countering of these sources of societal discord. It will be necessary to be cautious about the potential impact of AI systems that have destructive political and demographic biases (Suguri Motoki, Pinho Neto, and Rodrigues 2023). As humans increasingly rely on AI-generated content to make decisions, these systems will have an enormous influence to shape our perceptions and manipulate our behavior if left unchecked. Public-facing AI systems that exhibit fringe political bias will contribute to societal polarization, as users seeking confirmation bias may gravitate towards politically aligned systems, while those with different viewpoints may avoid them. Concernedly, it appears that GenAI will become a tool used to accomplish the *accelerationism* of societal disintegration.

Instead of being used by fringe elements to advance an agenda, AI systems can be directed to provide factual information on empirically verifiable issues. If these are based on legitimate elements, the content can offer diverse viewpoints and sources on contested topics that are often underdetermined. By doing so, these systems can help users gain insight, overcome in-group biases, and broaden their perspectives, potentially playing a useful role in defusing societal polarization. It is essential that language models claiming political neutrality and accuracy, like GPT-4 based models (OpenAI 2023), remain transparent about any biases they exhibit on normative questions, or to not be used by ideologues to drive a destructive narrative. It is the responsibility of the broader AI community to develop solutions such as we have explored in this paper, in order to address the developing dangers of GenAI for socially destructive actions. With the necessary CTR and control features it can be possible to ensure social well-being with methodical monitoring of GenAI and with pro-active countering of mal-info.

References

Abarna, S.; Sheeba, J.; Jayasrilakshmi, S.; and Devaneyan, S. P. 2022. Identification of cyber harassment and intention of target users on social media platforms. *Engineering applications of artificial intelligence*, 115: 105283.

Barbaro, F. 2022. *Analyse exploratoire et classification de textes*. Theses, Paris 1 - Panthéon-Sorbonne.

Barbaro, F.; and Skumanich, A. 2023. Addressing Socially Destructive Disinformation on the Web with Advanced AI Tools: Russia as a Case Study. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, 204–207. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394192.

Gabrielatos, C. 2018. *Keyness Analysis: nature, metrics and techniques*, 225–258. United Kingdom: Routledge. ISBN 9781138895782.

Giumetti, G. W.; and Kowalski, R. M. 2022. Cyberbullying via social media and well-being. *Current Opinion in Psychology*, 45: 101314.

Govers, J.; Feldman, P.; Dant, A.; and Patros, P. 2023. Down the Rabbit Hole: Detecting Online Extremism, Radicalisation, and Politicised Hate Speech. *ACM Comput. Surv.* Just Accepted.

Hardie, A. 2014. Log ratio: An informal introduction. *ESRC Centre for Corpus Approaches to Social Science (CASS)*, 1–2.

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Osman, K. K.; Barbaro, F.; and Skumanich, A. 2023. The potential for social media analysis to assess and optimize water management. In *9 th International Conference on Computational Social Science IC2S2*. Copenhagen, Denmark.

Peucker, M.; and Fisher, T. J. 2023. Mainstream media use for far-right mobilisation on the alt-tech online platform Gab. *Media, Culture & Society*, 45(2): 354–372.

Salunke, S. S. 2014. *Selenium Webdriver in Python: Learn with Examples*. North Charleston, SC, USA: CreateSpace Independent Publishing Platform, 1st edition. ISBN 1497337364.

Stocking, G.; Mitchell, A.; Matsa, K. E.; Widjaya, R.; Jurkowitz, M.; Ghosh, S.; Smith, A.; Naseer, S.; and Aubin, C. S. 2022. The Role of Alternative Social Media in the News and Information Environment. *Pew Research Center. URL https://www. pewresearch. org/journalism/2022/10/06/therole-of-alternative-social-media-in-the-news-andinformation-environment*.

Suguri Motoki, F.; Pinho Neto, V.; and Rodrigues, V. 2023. More Human than Human: Measuring ChatGPT Political Bias.