AboutUs: http://www.the-criterion.com/about/

Archive: http://www.the-criterion.com/archive/

ContactUs: http://www.the-criterion.com/contact/

EditorialBoard: http://www.the-criterion.com/editorial-board/

Submission: http://www.the-criterion.com/submission/

FAQ: http://www.the-criterion.com/fa/

Moderating Free Speech on Twitter

Klara Pakhomenko
RWTH Aachen University.

Article History: Submitted-06/07/2023, Revised-19/08/2023, Accepted-22/08/2023, Published-31/08/2023.

Abstract:

Social media acts as a powerful backbone for global communication and democracy. Twitter is a prime example, being one of the most wide-spread and influential forms of social media. The platform has gone through structural and functional changes in the hands of a new management that proclaims the protection of free speech as one of its main goals, which requires an analysis of what type of speech regulation, if any, is appropriate for the platform. This invites a utilitarian argument. In this essay, the radical utilitarian stance of John Stuart Mill on the value of free speech is presented, and Tweets, viewed as speech acts, are analysed in how much they can promote these values, as well as the harm they can bring. Finally, some potential guidelines for moderation are discussed.

Keywords: social media, Twitter, free speech, hate speech, speech acts, moderation.

1 Introduction

The rise of social media is one of the most fundamental breakthroughs within the history of human development. Global communication and democracy is fueled by it. Twitter is one of the most shining examples, representing a platform that can be accessed by virtually anyone, allows any variety of content, and can show it to any other user. Twitter not only accelerates and spreads out interaction in groundbreaking ways, it also transforms human speech into something entirely new, and allows for different kinds of interaction ruled by anonymity and targeted communication across the globe. With this shift, it undoubtedly unlocks a huge potential good for global discourse, but it also amplifies the harm that human speech is already used for, and adds brand new types of insidious danger.

In recent times, Twitter has gone through plenty structural and functional changes in the hands of new management, one which often puts a focus on free speech in its public communication about their goals for the platform. Thus, the question of balancing and regulating the massive possibilities presented by the platform against its

mine-field of hazards has never been more relevant. This type of question naturally invites a utilitarian argument, and in this essay, I will consider one of the most liberal utilitarian arguments for freedom of speech imaginable, that presented by John Stuart Mill in his work “On Liberty”. Mill’s ideas are often considered one of the most influential and radical accounts on the topic of free speech, and thus serve as an apt starting point for an analysis of Twitter, beginning at an extreme and then navigating the reality of the platform to find a suitable adaptation of Mill’s ideas.

I will first give a breakdown of Mill’s argument and how Twitter fits into his ideas on media and public spaces, as well as Mill’s greater body of work. Then, taking Mill’s argument as a starting point, I discuss the reality of Tweets as speech acts, their value in the context of free speech, and the harm they can cause, usual and novel. Finally, I propose a couple of fundamental guidelines for the moderation of free speech on Twitter, moving away from a blind copying of Mill’s radical ambitions into a more realistic and applied understanding of the platform as it is. It has to be noted that this essay is not, as it may appear, a line of argument contra Mill. Instead, I object against a flat interpretation of Mill that does not allow for the nuances and novelties of speech acts on Twitter. It would of course be convenient to simply cite Mill as a staunch defender of free speech, declare Twitter as an arena of free speech that has to be left unintruded, and go with a hands-off approach, as private companies often like to do. I aim to show that, even using Mill’s ideas on free speech as a basis, plenty of fundamental moderation on Twitter is justified and downright necessary.

2 Mill’s defense of free speech

John Stuart Mill’s seminal 1859 work “On Liberty” contains one of the most widespread defenses of free speech ever, often still cited as the classic line of argument in favor of a very liberal exchange of opinions (Mill 2021). In this section, I sum up his thoughts in their original form.

2.1 Mill’s line of argument

In the second chapter of “On Liberty”, Mill argues that censorship, no matter its intention, is always illegitimate and detrimental to the ones it aims to serve. When considering his reasoning, it must first be noted that the type of censorship that Mill considers takes shape as restrictions on speech enforced by the government or its people (Mill 2001, p. 18). This type of censorship can be enforced by law or by the

will of the general public, but not specific actors or groups. So, while not encroaching on the rights of private venues and publishers, Mill puts forth a very polemic argument for as few constraints as possible on what kind of speech should be allowed in public spaces by principle. When "On Liberty" was originally published, the most prominently considered public spaces were the press and what we would now call "mainstream media", but the argument can also be extended to spaces such as academia and, more generally, any large enough forum that allows for the trading of thoughts. Later in section 3, I will discuss how Twitter fits into this argument, but here I only reconstruct the classic case.

Mill takes a hard-line stance on censorship: "If all mankind minus one were of one opinion, and only one person were of the contrary opinion, mankind would be no more justified in silencing that one person than he, if he had the power, would be justified in silencing mankind" (Mill 2001, p. 18). To elucidate on this, Mill makes a case distinction. First, the stifled opinion could be true. If this is the case, then not only is humanity robbed of a true proposition, but everyone loses an opportunity to better themselves and update their belief system. (Mill 2001, pp. 19–22). Second, the silenced opinion could be false. Even then, it is not permissible to suppress it: If we want to understand our beliefs, we need to challenge them. Dealing with dissent serves as valuable training for any citizen and makes for a more sharp and well educated population. Even if you hold a true belief, it is not knowledge until you are able to argue for this belief and against its contra-position (Mill 2001, pp. 34–35). Furthermore, Mill argues that, in practice, this case distinction is a false one in the first place. Opinions generally do not have a binary value of true or false, but rather exist on a spectrum of truth, or there could be multiple, distinct truths, all of which should be heard (Mill 2001, p. 35). Finally, the decision to silence an opinion based on its truth or usefulness is not only detrimental but also illegitimate, as there are no infallible judges for either. (Mill 2001, pp. 23–24).

In short, every statement deserves to be put forth into the shared public discourse. No proclamation of thoughts and ideas should ever be hindered by the government or the people, as a free expression of opinion will always be advantageous to the whole of society. This is the most pure and unrestrained form of Mill's dream. In the following, this dream will be put into the greater context of "On Liberty", extended to a modern understanding of public spaces, and tested against the manifold reality of Twitter to see what remains.

2.2 The Harm Principle

Mill's stance on censorship is an uncompromising one and serves to outline his guiding principle for a just and healthy public space, but it comes with an asterisk within his own work. This is the simple but effective Harm Principle, which acts as a golden rule for the whole of "On Liberty": "That principle is, that the sole end for which mankind are warranted, individually or collectively, in interfering with the liberty of action of any of their number, is self-protection. That the only purpose for which power can be rightfully exercised over any member of a civilised community, against his will, is to prevent harm to others" (Mill 2001, p. 6). In fact, proposing this principle and inspecting its various implications on society is the general aim of "On Liberty", and his thoughts on the freedom of speech are only part of this bigger project. As with all of Mill's practical philosophy, this is motivated by his utilitarian ideals (Macleod 2020).

So, having placed Mill's advocacy for free speech within his general body of work, it becomes clear that Mill argues with a utilitarian aim, and thus not for completely unrestrained free speech, but for a public discourse where any opinion is able to be voiced, as long as the act of voicing it causes no harm to others. At this point, it becomes clear that it will be necessary to pin down what exactly "harm to others" entails, as well as "the act of voicing" an opinion. Both of these aspects must be investigated and the application of the Harm Principle to the case of Twitter will hinge on this distinction. Keeping in mind Mill's very liberal ambitions and his aim to have as few restrictions as possible, it is most apt to consider a very narrow interpretation for this essay, and focus on cases of direct harm which are easily provable. In section 4, this pressing question will be inspected further, with an eye on Tweets as acts of speech rather than simply propositions.

3 Mill and Twitter

If we wish to inspect Twitter through the lens of Mill, we first have to make sure whether the type of censorship that Mill argues against can be reasonably compared to censorship on Twitter, as the structure of public discourse on social media now is in many ways completely different to the kind that Mill had in mind.

3.1 Is Mill's argument applicable to Twitter?

Recall that the type of censorship that Mill considers is restrictions on speech enforced by the government or the general public (Mill 2001, p. 18). But, can this even apply to Twitter? On a surface level, Twitter is a private company that does not work in service of any government. Neither is it a platform controlled by its users, but rather a select group of developers and decision makers. Given this understanding of the platform, it might appear as if Twitter shouldn't be held to the standard that Mill lays out. But to take this view is to completely underestimate the epistemic power that social media giants hold. Should Twitter decide to ban the use of certain terms or the expression of certain opinions on its platform, then this no doubt has large scale effects on public discourse, comparable to a governmental restriction of the press, or perhaps even going beyond what a government can do. After all, social media platforms are international and have low barriers of entry, and thus affect a much wider mass of citizens than any single government.

In fact, the existence of social media can be considered a manifestation of Mill's dream: Anyone with an internet connection may upload their thoughts on any subject directly into a public forum spanning essentially the entire world. A more liberal exchange of information is very difficult to imagine. The amount of speech facilitated by social media giants such as Twitter is unprecedented throughout all of history, and the wealth of true propositions and intellectual discourse generated from these platforms is priceless in its utility. Taking this into account, it only serves Mill's utilitarian aims to treat Twitter the same as one would a government or a population of people, as the global discourse rises and falls with the rules put in place by companies like it. At this point, I draw the tentative thesis that perhaps, following Mill's line of argument, one should make as much use as possible of this wonderful well of utility and censor only those tweets that are in clear violation of the Harm Principle, while everything else is allowed. After all, as long as no existing laws are broken and no rights are invaded, the potential for discourse and free exchange of information is so great that nothing should stand in its way. To see if this thesis holds, one must understand what censorship on Twitter may look like, and what kind of harm Twitter creates, to be able to determine the cases where intervention is necessary and legitimate, and where it isn't.

3.2 What is censorship on Twitter?

Twitter differs from the cases considered by Mill not only regarding who's in control, but even the structure of discourse itself is also categorically different due to mass interpersonal communication, as well as filtering (see for example Ulen 2001, Nguyen 2020). Instead of users just opting in or out of set news-feeds like one would turning on the radio or reading a newspaper, the spread of content is controlled and accessed by every user, as well as an algorithm that influences who sees what when. User experiences are highly personalized and there are many dimensions to how a user is heard by the public. This muddies the idea of what can be considered censorship: Is it censorship if a user is allowed to use the platform freely, but gets deliberately ranked lower by the algorithm and isn't shown to new users in the same way as others? Is a user hindered from voicing their opinion if their account is publicly accessible, but every tweet comes with a big red alert marking it as untrustworthy? Does simply inconveniencing a user from accessing content fall under the kinds of restriction that Mill so convincingly argues against? Here, all the weight lies on what is considered hindering a user from being heard by the public.

Considering one extreme, one might want to call all of the above cases censor-ship, and attempt to imagine the ideal Twitter as a social media platform where all opinions are heard equally. However this is already not the case, and would require far too radical structural change. To consider one example, mixing entertainment or shock value into posts amplifies engagement and the algorithm will show the most engaging tweets to the most people, ultimately maximizing the time users spend on the app and thus company profits. This is what private companies do. In this way, the opinions of "boring" users are in some way suppressed by the platform itself, discouraging equal and temperate discussion. However, this is in no way a new phenomenon introduced by Twitter, and has already been the way of public discourse during the publication of "On Liberty". Opinions do not reach the public based on their content alone, but also the resources that the speaker has access to: Nepotism, financial status, rhetoric, character and appearance, and so on. Uprooting this basic state of things using social media might not be possible. To imagine a platform where everyone is being heard absolutely equally, on the scale of Twitter, is a fruitless pursuit within the current structure of society. So instead, I will focus on the opposite end of the spectrum in this essay and consider only the clear cut cases, specifically deleting tweets or banning a user from all participation. This would be akin to the

government forcibly removing contro-versial articles or authors out of the shared public space. These strong cases of intervention of course need strong arguments in support of them, but when one considers the reality of Tweets as speech acts and the novel mechanisms of Twitter amplifying their reach and insidiousness, several areas of concern open up.

4 The role of Tweets

What is a Tweet really? In order to understand the potential harm to others caused by Twitter posts, one needs to pin down what role speech plays in Mill's argument, and how Tweets fit into this idea, as well as where they differ.

4.1 The role of speech in "On Liberty"

When Mill so vigorously upholds the value of speech and argues in favor of free discourse in line with his utilitarian ideals, he primarily means an exchange of information. In Mill's original line of argument, speech is first and foremost propositional. A statement can have true or false components, it can have a use for the project of science and the pursuit of human development. Still a held opinion can not, in itself, be harmful. There is no mention across the entire second chapter of "On Liberty" of speech used to accomplish a goal, said within a relevant relation to some recipient, or generally speech doing anything worse than simply being wrong. Speech doing something is where harm to others starts coming into question.

Chapter two of "On Liberty" is concerned with forming an opinion but not yet acting on it, and to tweet means to act. Therefore, we have to turn to the much more difficult topic of speech as an action. Mill himself notes this when he writes "even opinions lose their immunity when the circumstances in which they are expressed are such as to constitute their expression a positive instigation to some mischievous act" (Mill 2001, p. 52). He brings up the example of a corn dealer: Accusing the corn industry of robbing the poor is, in itself, not problematic it does however become an issue if one states this to an angry mob ready to break down the doors of a corn dealers house. This implies a rather direct and limited interpretation of the Harm Principle, and it is vital to keep this in mind when applying Mill's ideas to Twitter. In order to do this, the propositional element of a Tweet and the action of tweeting it have to be clearly distinguished. And furthermore, the action of tweeting has to be better understood.

4.2 Speech Acts

At this point, it is useful to give a short and general anatomy of speech acts in order to carry out a solid analysis of Tweets. John L. Austin, one of the most influential language philosophers of the 20th century, describes a way in which speech acts can be understood by breaking them down into the locutionary, illocutionary, and perlocutionary act.

> Take for example the speech act of saying "I think this is terrible!" to a student upon reading their essay. The locutionary act is the utterance of the phrase itself, as well communicating the content of a phrase, that is saying the sentence out loud and referring to yourself as "I" and the student's essay with "this" (Austin 1962, pp. 94–98). The illocutionary act is not just the uttering of content, but also the force of the utterance, which is dependant on the way something is said as well as its context and speaker (Austin 1962, pp. 98–101). When said to a student by their professor after the work was handed in, the phrase in question is a criticism and also an announcement that there will be an unfavourable grading. If the same thing is said to the student by a colleague, this might still be a rude criticism but it is not said with any authority, and does not come with a warning of consequences. Thus, these are two different illocutionary acts. Finally, the perlocutionary act is the actual effect on the thoughts, feelings, and actions of the speaker or audience (Austin 1962, p. 101). Here, this would be the act of hurting the students feelings, or informing them that they will fail the class.

This is of course only a brief overview of speech acts, but it does indicate what parts of tweets we are interested in. The locutionary element of a Tweet might be of great interest for the philosophy of language and grammar and narrowing down what exact proposition has been made, but the actual effect and possible harm caused by a Tweet is in its illocutionary and perlocutionary force. Since this essay is concerned with the moderation of Twitter and thus Tweets as speech acts, I am going to be concerned mostly with the illocutionary and perlocutionary force. These forces are determined by the context of an utterance, and Twitter presents a radical transformation of the environment of speech acts, as well as new ways to use it. In the following, I am interested in exactly this aspect of Tweets.

4.3 The role of speech on Twitter

At long last, we move to the central question: Considering Tweets as speech acts, how does Twitter allow Mill's dream to flourish, and where does it open up new areas of concern?

4.3.1 Understanding Tweets as speech acts

To interpret Tweets as speech acts, we must ask: What is being said on Twitter? How is it being said? Who is the speaker, recipient, and audience? The latter questions can be answered simply: In a Tweet, the author of the post speaks to potentially every user on the platform. The recipient is first and foremost the author's followers, but the audience extends to every other user, as well as visitors on the rest of the internet that are not logged in. And regarding the first question, anything goes. Twitter is what happens when one puts news, political debate, friendship, and porn into the same box. There is no real limit to the type of content on the platform, with any real life interaction having a counterpart, as well as new ways of interaction being possible.

This blurring of the lines between types of speech is a defining feature of Twitter and is precisely what makes it such a monolithic platform, but it also presents a problem for the potential utility of discourse on the platform. The short format of Tweets, as well as the competition for attention built into the platform, encourages users to not actually engage in any kind of thorough discussion - that is, give context, cite sources, and develop nuanced points of view - but instead sell their opinion with the most snappy and instantly affecting post possible. When you're still within the original context of Mill's argument for free speech, one might say this type of post, and even statements in bad faith, are important because they bring opinions into public consciousness and force others to reflect on it and possibly argue against it. Yet even if this were attempted, arguing against an opinion on Twitter leads the respondent into the same trap. In order to survive and be heard, users must turn their counter-argument into another boiled-down show of entertainment, and even if the proposition at the Tweet's core might be valuable, it can not reach the desired effect in its perlocutionary act due to the context it's being forced through.

This sheds some light on how things are being said on Twitter. Due to its format and general structure, the illocutionary act of a Tweet within a debate is partly an attempt to convince others, but importantly also a ploy to pull those users already

on your side in and generate engagement, so that, in its perlocution-ary act, your side may win the number game and create a feeling of superiority within the part of the audience that agrees with you. After all, carrying out a full argument including support for your claims on Twitter is a strict impossibility due to its format, if the resulting chain of Tweets is meant to be seen and processed by its intended audience. Thus, highly immediate and shareable "dunks" on the opposition is the type of speech act that will be most commonly found on Twitter debates, not as an unfortunate misuse but due to the very structure of the plat-form. This effect of users that already agree being grouped together is visible in the very real echo chambers on Twitter (Cinelli et al. 2021). It also muddies the idea that the audience of a Tweet is all of Twitter - in reality, for many users, their audience might be only a bubble within it.

All of these considerations don't spell doom for Mill's idea of free speech right away, but they do limit it. Outside of the context of a debate, one might see bundles of information going around to spread awareness, often together with links to charity organizations in times of political and social trouble. As before, despite its indisputable positive utility, this is again the type of speech act that will reach those who already agree to the given cause. This leaves all the other types of Tweet, which might generate entertainment, friendship, or other positives, but don't have much propositional value as speech acts. All in all, considering the reality of speech acts on Twitter, the potential for utility generated from free speech on it, and thus the potential fulfillment of Mill's dream, seems rather grim. This much can be said about Tweets as speech acts and how they can help to achieve a fruitful global debate on matters of opinion. It remains to ask: What about the harm they can bring?

4.3.2 Harmful speech carrying over to Twitter

A lot has already been said about harmful speech offline and its interaction with Mill's argument. Twitter mirrors and amplifies the harm caused by "usual" speech, and thus re-ignites these debates. In this essay, I mostly focus on the novel sit-uations created by Twitter and therefore can not dive into each of these topics and their transformation when introduced to the platform. However, they may absolutely not be left off the table! It will have to suffice to name a few that are vital to consider.

The most immediately apparent case is hate speech. In The Harm in Hate Speech, Jeremy Waldron focuses on the social effect of hateful speech and argues that, in a society where human dignity is considered a basic right, it is not permis-sible to

express hateful thoughts with conduct that insults this dignity. This is an extension of the Harm Principle in its original form, as the harm caused by hate speech in its perlocutionary act is not direct and easily provable, but this idea is in line with what many governments already do. Twitter allows hate speech to spread in much quicker and more effective ways, especially when it's slid into posts that are entertaining and quick to catch on, and therefore Waldron's arguments need to be re-examined and can be strengthened for the case of Twitter.

Going even further, Joel Feinberg suggests the "Offense Principle" in his work Offense to Others, which states that not just harm, but mere offense is already a legitimate reason for government intervention on speech. The restrictions should of course be less severe than when speech causes harm, but should nevertheless be considered. On Twitter, the potential to reach and offend audiences is of course much greater. In fact, shocking and disgusting content is in some ways encouraged by the algorithm, as engagement, positive or negative, will always cause a post to become more successful. Simultaneously, Twitter gives users the ability to single out specific audiences and hand-tailor offensive content especially for them. So in the context of Twitter, Feinberg's ideas have to be re-considered, and if they have any legitimacy in the world of mainstream media, they become even more pressing for Twitter.

On the topic of offense, the most extreme example may very well be the usage of slurs. Slurs no doubt fall into the category of speech Feinberg speaks of, as generally audiences are often offended by slurs and have every right to be, as Ren´ee Bolinger discusses in "The Pragmatics of Slurs". There are arguments on prohibiting the usage of slurs due to the violation of respect between speakers in the illocutionary act of slurs, and Twitter allows the targeted selection of slurs based on a user's public post history and profile description, exacerbating the issue. It is doubtful that such expression should even be defended by Mill's argument, as a slur can be exchanged for different words with the propositional value of a speech act remaining equivalent, but the illocutionary and perlocutionary act that come with it being much more apt for public discussion. This should cover some of the "regular" harmful speech carrying over to Twitter, but Twitter also allows for entirely new phenomena to emerge.

4.3.3 Novel harm created by Twitter

Like all social media, Twitter massively restructures the way that people interact. The veil of anonymity on the one hand, as well as the possibility of targeting groups of users across the globe on the other, allows for new types of interactions.

The possibly most classic example is cyberbullying. Cyberbullying is of course not exclusive to Twitter, but represents a serious problem with wide reaching consequences on many platforms, and the harm that can be caused by it is very real and very direct (Cohen-Almagor 2020). A lot of attention has already been paid to the cases where real life interactions spill over onto social media and use anonymity and constant barrages of posts to cause more harm, but cyberbullying does not have to have its roots in real life conflict. Cases of bullying users into potentially life-threatening situations can start and end online, and users can easily make themselves more vulnerable on Twitter by publicly signaling their status as a minority or otherwise giving bullies easy ways in. There should be no doubt that it is the responsibility of platform owners and regulators to reduce the risk of this type of harassment, as has been argued many times before.

It might not be a brand new observation that social media is host to a vast variety of “cybercrimes”, see Crime, Justice and Social Media for an overview on some. Most importantly though, Twitter also separates itself from other social media and allows for something more insidious. Think about the phenomenon of dogwhistles. A dogwhistle can be defined as a speech act masquerading as a reasonable part of discourse, but deliberately allowing for a second, private inter-pretation, that speaks only to a pre-selected audience and conveys a coded message (Saul 2018). There are plentiful types of dogwhistles, but the important part for the case of Twitter is the use of one set of words to really talk about another. The illocutionary act of a dogwhistle is shaped by this intention to mislead a public audience and simultaneously mobilize a smaller, targeted group. A classic example of a dogwhistle would be raising discussion about urban crime and showing cer-tain statistics about demographics, while really trying to push their claims about the danger that black people apparently pose to peaceful society. This is excel-lent propaganda. A general audience unaware of the speaker’s intentions will nod along and hear them out, taking their arguments at face value and amplifying their reach, while the racists will nod along and spread the message equally, knowing that the speaker is on their side and agreeing with the second, coded message.

By themselves, dogwhistles are an especially sneaky form of hate speech and should be part of the discussion on which Tweets are harmful in the "harm against dignity" sense, or otherwise offensive. On Twitter, the possibility to infuse dogwhistle-like speech into the public arena is especially enormous. This is due to the user's access to Hashtags, and more generally trending words. Often, investi-gating a Hashtag or a trending phrase will reveal a whole plethora of Tweets one might not have naively expected to see - click on "Free Speech", and you will see an army of white users proclaiming their right to use slurs like the "n-word", or click on "#MichaelJackson" and instead of stories about the king of pop, you are bom-barded with complaints about apparent "cancel culture", and how celebrity child abusers shouldn't be scrutinized by the public because it is none of the public's business. There is even the possibility to hijack a trend that is already circulating, when something neutral gains attention and then allows for the mobilization of users that wish to steer the discourse on Twitter into a direction that suits them.

To illustrate the nature of the dogwhistle-trend-hashtag with something more local to Twitter, consider logging in and checking your trending tab to find the user-base discussing safety in public bathrooms, and the legality of sports competi-tions. Innocuous at first glance, these trends are almost without exception started by bigoted users that would like to regulate transgender people out of existence. Bathroom safety is really about the claim that trans people use their gender iden-tity in order to get access to a different bathroom, because sexually assaulting "the opposite sex" is easier this way. When sports competition rules trend on Twitter, this is really about the fact that trans women are accused of having too much of an advantage within their gendered space and thus should not be allowed in sports, or better not anywhere. To go into the legitimacy of these arguments is besides the point. The relevant part here is how the structure of Twitter can be used to hijack or create platform-wide signals that attract the attention of like-minded individuals who spread covertly hateful conduct by optically innocent means, and simultaneously lure in swaths of users that don't know any better and get exposed to dubious arguments made with the intention of lowering the social status of a targeted group, eventually hopefully taking away their rights. This isn't about truth-seeking or debate at all. It's about the mobilization of hate groups across the globe and sneaking propaganda into faux-argumentative posts, or even just entertainment. Even posts pretending to be funny quips or viral images can carry second, coded messages, targeting specific groups of users. Through

this method, Tweets often function not as carriers for worthy exchanges of information and argument, but as rather potentially very harmful and effective weapons.

5 Moderating Twitter

Twitter has a massive potential for supporting free discourse on any topic, but like any public forum, there need to be rules of conduct to allow for this potential to unfold, and to keep the platform from becoming host to a plethora of harmful practices. In section 3, I proposed that, in light of Mill's arguments and the huge potential of Twitter to align with his ideas, Twitter should be moderated as little as possible, only removing posts or banning users for clear violations of the Harm Principle. I also said that strong cases of intervention need strong arguments to support them. In section 4, I brought up several of these arguments. I analyzed the reality of Tweets as speech acts and their limited use for debate, and went through a whole array of cases where either the Harm Principle has been violated by Tweets, or it appears that the cost of certain Tweets is so great that it might be necessary to extend the Harm Principle to capture them. It remains to bring everything together and see what kind of moderation should be fit for Twitter.

To this end, I reiterate that it seems very difficult to see Twitter as a bastion of free speech in the way that Mill might have imagined an ideal public space. Other social media with different structures and moderation techniques might be more suited for these goals, but as I discussed, the propositional value of arguments on Twitter is limited by the very foundations of the platform. Therefore, I argue that, since Mill works within a utilitarian context, freedom of speech on Twitter can not be protected with the same vigilance as other spaces by Mill's arguments. The potential for the spread of hate speech needs to be reconsidered due to the virality of opinions on Twitter, and if the possibility of dignified and restrained discussion is to be upheld, the most extreme cases of hate speech, especially including the use of slurs, needs to be kept to a minimum, and offending Tweets as well as users that engage in this behaviour consistently need to be taken off of the platform. Their contribution to discourse is minimal, and the damage done to other users as well as the given discussion can be severe.

This stance only reflects the need for rules of conduct in any public space, which is not contra Mill, as it is a requirement for proper discussion to be possible in the first place. If the Harm Principle does not capture the worst cases, it needs to be

amended or supplied with rules for how speech is to be expressed in a fruitful conversation, regarding the illocutionary and perlocutionary act of speech, not what is expressed propositionally. And as for the cases of cyberbullying and other cybercrime, the original and most strict version of the Harm Principle already implies the need for intervention, as the rights and personal space of single users gets invaded and direct harm can be observed. Here too, repeatedly offending users and the Tweets in question have no place on Twitter.

All of this being said, one may not be too hasty in considering all "hateful Tweets" as irrelevant and harmful speech to be removed without question, as even bad faith arguments need to be discussed and disproven and not simply thrown away, if the truth is to be maintained diligently and held with proper care. There has to be a careful distinction of hateful conduct, and keeping in mind the Offense Principle, a way to single out especially offending users and reacting accordingly: banning users that are a net negative on public discourse. These rules and sanctions are only a necessary part of cultivating the right kind of debate to achieve Mill's dream, and don't in themselves stand in the way of valuable free speech.

And considering the Twitter-specific issue of dogwhistles turned into trends and Hashtags, special care has to be taken. I propose identifying and investigating new trending movements, keeping tabs on signals started by hateful users in covert attempts to sneak their ideas into public consciousness, and supplying the corresponding trending pages with information on what the trend is, possible second coded interpretations, and links to reputable sources on the topic. This needs constant monitoring and research, but on a platform as wide-reaching as Twitter, this appears unavoidable. The seeming absence of any such caution is quite troubling.

I wish to make a final point on the topic, which is this. The general state of Twitter is not brought about by any single one of the issues discussed in this essay. It is not the failure of the platform to support productive speech, or any variety of hate speech, or cybercrime, or Trojan horse trends by themselves. It's all of them acting in concert, causing a pervasive miasma on the platform as a whole. To consider a final example, think about the events of January 6th in America, which not only caused harm to many citizens lives, property, and dignity, but also historically unprecedented damage to American democracy. Even if it is possible to imagine an uprising of the people in agreeance with Mill's philosophy when it comes to the protection of certain rights, there was no such justification for Janurary 6th. No argument by Mill could

support such acts against basic liberal values. Twitter was no doubt a key component for this, of course allowing the involvement of the American president on the platform, but also previous actions by a whole web of users allowing the conditions to fester and eventually snap into place. Trying to pin down what happened to some kind of singular root cause is virtually impossible due to the amount of actors and mechanisms behind it. So instead, when trying to capture a threat like this, one needs to cast a wide net. If we want to save Mill's dream and have it live on on Twitter, we need to broaden our view to see the entire picture, and extend the Harm Principle to Twitter in the necessary ways.

6 Conclusion

As with many revolutionary inventions of mankind, Twitter is a double edged sword. And like any deadly weapon, it has to be handled with grace and precision, with its effects on the world fully understood. John Stuart Mill's ideas on free speech give a hopeful view of how invaluable speech is when it is freely circulated, and at first glance, Twitter might seem like the perfect candidate to materialize this vision - but, as I have discussed at length, the reality of Tweets as speech acts tells a different story. Due to its structure and incentives, Twitter is often not an appropriate space for valuable discussion, and the type of harm that it can create often challenges the positives proclaimed by Mill. Thus, in a utilitarian context, one is forced to devise ground rules for moderation on Twitter, in order to limit extreme cases of hate speech, get malicious users out of the picture, curb the possibility of cybercrime, and most importantly, work against the misleading and harmful nature of certain trends and social movements on the platform that get broadcasted to the entire userbase.

The current development of Twitter in the hands of new management runs against the ideas presented in this essay in two ways. First, it is unlikely that Twitter is attempting to follow the radical stance on free speech Mill has outlined, despite owner and CEO Elon Musk's public display of commitment to allowing all legal speech. Musk has personally banned prominent journalists circulating public information on his whereabouts and reporting on it, citing a case of "doxxing" where it is not applicable, in this case actively working against free speech on Twitter (Darcy 2022); (Clark, Heath, and Lopatto 2022). On the other hand, he has also discarded any consideration of the Harm Principle or its extensions by reinstating Donald Trump's

account after his involvement in the January 6th riot on the platform, and undoing policies meant to lessen the amount of intentional misinformation on Covid-19 and election fraud (Thompson 2022). It is also now possible for users to buy verification for themselves and more easily misinform and mislead unknowing targets when engaging in the type of dogwhistling I have described in this essay.

A development of the Twitter platform aligned with liberal ideas on free speech will thus have to change in two ways. First, an actual commitment to allowing as much speech as possible has to be realized, and second, the moderation rules have to incorporate and extend the Harm Principle in order to allow free speech to flourish. A laissez-faire "marketplace of ideas" attitude in order to protect and uphold the value of free speech in the way that Mill imagines is an option for many platforms, but when considering specific cases, there need to be specific rules of conduct in order to enable fruitful discourse and keep it from causing unprecedented damage to society. And in the case of Twitter, the scales are unfortunately tipped towards more moderation.

Works Cited:

Austin, John Langshaw (1962). How to Do Things with Words. Clarendon Press. Bolinger, Ren´ee Jorgensen (2017). "The Pragmatics of Slurs". In: Noˆus 51.3, pp. 439–462. doi: https://doi.org/10.1111/nous.12090. eprint: https: //onlinelibrary.wiley.com/doi/pdf/10.1111/nous.12090. url: https: //onlinelibrary.wiley.com/doi/abs/10.1111/nous.12090.

Cinelli, Matteo et al. (Mar. 2021). "The echo chamber effect on social media". In: Proceedings of the National Academy of Sciences 118, e2023301118. doi: 10.1073/pnas.2023301118.

Clark, Mitchell, Alex Heath, and Elizabeth Lopatto (2022). Elon Musk starts banning critical journalists from Twitter. url: https : / / www . theverge . com / 2022 / 12 / 15 / 23512004 / elon - musk - starts - banning - critical - journalists-from-twitter (visited on 03/30/2023).

Cohen-Almagor, Raphael (July 2020). "Moral Responsibility and Social Network-ing: Cyberbullying and Lessons from the Megan Meier Tragedy". In: European journal of analytic philosophy 16, pp. 75–97. doi: 10.31820/ejap.16.1.4.

Darcy, Oliver (2022). Elon Musk bans several prominent journalists from Twitter, calling into question his commitment to free speech. url: https://edition. cnn . com

/ 2022 / 12 / 15 / media / twitter - musk - journalists - hnk - intl / index.html (visited on 03/30/2023).

Feinberg, Joel (1984). Offense to Others. Oxford University Press USA. Macleod, Christopher (2020). “John Stuart Mill”. In: The Stanford Encyclopedia of Philosophy. Ed. by Edward N. Zalta. Summer 2020. Metaphysics Research Lab, Stanford University.

Mill, David van (2021). “Freedom of Speech”. In: The Stanford Encyclopedia of Philosophy. Ed. by Edward N. Zalta. Spring 2021. Metaphysics Research Lab, Stanford University.

Mill, John Stuart (2001). On liberty. English. Kitchener, Ont.: Batoche Books.

Nguyen, C. Thi (2020). “ECHO CHAMBERS AND EPISTEMIC BUBBLES”. In: Episteme 17.2, 141–161. doi: 10.1017/epi.2018.32.

Salter, Michael (Jan. 2017). Crime, Justice and Social Media. isbn: 978-1138919679. doi: 10.4324/9781315687742.

Saul, Jennifer (Aug. 2018). “Dogwhistles, political manipulation, and philosophy of language”. In: pp. 360–383. doi: 10.1093/oso/9780198738831.003.0013.

Thompson, Stuart A. (2022). Musk Lifted Bans for Thousands on Twitter. Here’s What They’re Tweeting. url: https : / / www . nytimes . com / 2022 / 12 / 22 / technology/musk-twitter-bans.html (visited on 03/30/2023).

Ulen, Thomas (Oct. 2001). “Democracy and the Internet: Cass R. Sunstein, Republic.Com. Princeton, NJ. Princeton University Press. Pp. 224. 2001”. In: SSRN Electronic Journal. doi: 10.2139/ssrn.286293.

Waldron, Jeremy (2012). The Harm in Hate Speech. Harvard University Press. isbn: 9780674065895. url: http://www.jstor.org/stable/j.ctt2jbrjd (visited on 08/21/2022).