



# Journal of Media Ethics

## Exploring Questions of Media Morality

ISSN: (Print) (Online) Journal homepage: <https://www.tandfonline.com/loi/hmme21>

# Are You Sure You Want to View This Community? Exploring the Ethics of Reddit's Quarantine Practice

Caitlin Ring Carlson & Luc S. Cousineau

To cite this article: Caitlin Ring Carlson & Luc S. Cousineau (2020) Are You Sure You Want to View This Community? Exploring the Ethics of Reddit's Quarantine Practice, Journal of Media Ethics, 35:4, 202-213, DOI: [10.1080/23736992.2020.1819285](https://doi.org/10.1080/23736992.2020.1819285)

To link to this article: <https://doi.org/10.1080/23736992.2020.1819285>



Published online: 11 Sep 2020.



Submit your article to this journal [↗](#)



Article views: 53





View related articles [↗](#)



View Crossmark data [↗](#)



## Are You Sure You Want to View This Community? Exploring the Ethics of Reddit's Quarantine Practice

Caitlin Ring Carlson <sup>a</sup> and Luc S. Cousineau <sup>b</sup>

<sup>a</sup>Communication Department, Seattle University, Washington, United States; <sup>b</sup>Department of Recreation and Leisure Studies, University of Waterloo, Ontario, Canada

### ABSTRACT

In the United States, social media organizations are not legally liable for what users do or say on their platforms and are free to regulate expression in any way they see fit. As a result, dark corners of the Internet have emerged to foster communities whose sole purpose is to create and share content that subjugates members of traditionally marginalized groups. The subreddit, */r/TheRedPill*, is one such community. This article explores whether hiding this offensive content through digital “quarantine” or removing the community altogether is more ethically justifiable. We draw on theorizing about the ethics of social media content moderation to develop a framework for ethical decision-making based on transparency, corporate social responsibility, and human dignity to guide decisions about content removal. Using */r/TheRedPill* as a case study, we argue that the most ethically justified course of action is for Reddit to remove the site entirely from its platform.

### ARTICLE HISTORY

Received 3 August 2020  
Accepted 1 September 2020

Let's start with the obvious. American women like to be choked, spanked and have their faces spit into. They basically want the French Nazi collaborator treatment post WW2. Clearly, not every woman's idea of a romantic evening is to be covered in piss and pelted with rotten vegetables while a crowd watches and yells shame! I get that, and I'm not making a totalizing claim. Not all women, mostly the liberal ones.

What is obvious however is that a woman who voted for Trump is less likely to enjoy a romantic evening of peepee-poopoo on her face than a woman who voted for Hillary Clinton. There's a very simple explanation for this. Women who voted for Trump voted for a big protective wall to keep them safe from dirty Mexico. These are the type of women who protect themselves with rules, standards and limits. Unsurprisingly, Trump voting women are more religious because the Bible is basically just a big book of rules.

The above are the first two paragraphs from a September 19, 2019, post on the popular media sharing and discussion website Reddit.com. Specifically, this post comes from a senior contributor and well-known participant of the Reddit sub-community (called a “subreddit”) */r/TheRedPill* – a community dedicated to “discussion of sexual strategy in a culture increasingly lacking a positive identity for men” (*/r/TheRedPill*, 2019). *TheRedPill* is one part of a larger, albeit loose, collective of men's groups called *The Manosphere* (Ging, 2019; Gotell & Dutton, 2016), which include well-known communities such as Incels (Reeve, 2018) and lesser-known groups such as MGTOW (Men Going Their Own Way) (Futrelle, 2011). The tagline of */r/TheRedPill*, as well as the content of this post, are illustrative of much of the discussion in this subreddit and provide windows into the underlying ideological and sex-gender positioning of the community and its membership. */r/TheRedPill* had more than 400,000 subscribers (as of October 19, 2019 – the last time numbers were published for this now

**CONTACT** Caitlin Ring Carlson  [carlso42@seattleu.edu](mailto:carlso42@seattleu.edu)  Seattle University, Seattle, WA 98122

This article has been republished with minor changes. These changes do not impact the academic content of the article.

quarantined subreddit), and this post has been commented on and voted on hundreds of times, and likely read many more.

TheRedPill community is not the only one with contentious or troubling content on Reddit. The website, a strong supporter of free speech since its inception, has been (and is) home to a number of communities with controversial views, including */r/the\_donald*, */r/KotakuInAction*, */r/fatpeoplehate*, and */r/beatingwomen*.<sup>1</sup> Reddit allowed content on the site to be (mostly) free of censorship until 2015 when, under the short-lived leadership of Ellen Pao, it introduced its first widely actionable harassment policy, and with it a series of sanctions on users and subreddits (Moreno, Pao, & Ohanian, 2015). Users have been banned, and subreddits have been warned, limited, quarantined, and in some cases, banned and deleted from the site. Quarantine is of particular interest as it occupies a middle ground where a sub-community continues to operate but is secluded from the rest of Reddit and its users. When a subreddit is quarantined, it remains active on the site, but does not appear in aggregated content for non-community users, on */r/all*, and has an additional layer of warning before users can enter the community.

In this article, we use the */r/redpill* subreddit as a lever to open a discussion about the ethics of content moderation by private companies such as Reddit Inc. and how they might address issues of free speech, harassment, and other troubling behaviors on their platforms. To answer this question, we introduce Reddit as a website and service and explore the legal framework that sets boundaries for U.S. social media companies. We then provide an overview of the process of content moderation as it is applied broadly by social media companies in general, and Reddit in particular. Next, we explore existing ethical frameworks for social media content moderation such as the Santa Clara Principles (2020), the corporate social responsibility model (Carroll, 2016), Bowen's (2013) proposed deontological framework, and Rawls' (1971) veil of ignorance and use those to develop a decision-making model for social media organizations to employ as they navigate questions about content removal. Finally, we apply the proposed framework to determine whether quarantine represents an ethical approach to dealing with the volumes of hateful and harassing content that appear on social media and Reddit in particular.

## About Reddit.com

Reddit is a user-contributed, aggregated media site that “bridges communities and individuals with ideas, the latest digital trends, and breaking news (... okay, and maybe cats)” (reddit.com, 2015). Reddit has three mainstays of functionality: (1) The ability for any user to share content; either original text, image, or video posts, or content brought to the platform from other sites; (2) The ability for users to vote on that content; the voting system is a logarithmically weighted positive and negative attribution measure which accumulates over time for individual posts and for users in the form of valueless internet points, or “Karma”; and (3) user discussion on those posts, where discussion takes the form of asynchronous Bulletin Board System (BBS)-style postings and where users can comment on posts, and reply to comments, creating long and sometimes complex message threads. The site is divided into “subreddits” that allow for aggregation of content by subject (e.g. *reddit.com/r/gaming/as* a clearinghouse for game-related content, or *reddit.com/r/food/for* discussions and photos of food), with each subreddit consisting of content submitted by users and moderated by volunteer moderators from that user community.

Users are able to create their own subreddits, and as of July 2020 there were 2,252,647 subreddits with an average of 1,950 new subreddits created per day (Reddit Metrics, 2020). Users are able to subscribe to individual subreddits, which are aggregated onto a “front page” displaying the most popular posts from their subscriptions. Non-users may also explore Reddit, but without the individualized front-page experience. Visiting users default to the */r/all* page, which aggregates the most

<sup>1</sup>*/r/the\_donald*, */r/fatpeoplehate*, and */r/beatingwomen* have all been banned from the site (*/r/the\_donald* in June 2020). */r/KotakuInAction* remains active and open on the site.

popular content across most subreddits (some subreddits with adult or contentious content (including quarantined subreddits) are excluded from the aggregation algorithm of/r/all/), and visitors are not able to vote on posts or add comments.

Although there are some rules concerning posts (e.g., no child pornography), until 2015 Reddit Inc. maintained a broad and ambiguous policy of “supporting free speech,” and did little policing of content, following its mission to “help people discover places where they can be their true selves, and empower our community to flourish” (Reddit.com, 2016). The radical openness to most content and kinds of speech was essential in the site’s founding vision and was a vital factor for the site’s early success, according to its founders (Ohanian, 2016). This position has left Reddit reluctant to remove or censor content posted to the site, although moderators are free to do so (Marwick, 2017; Massanari, 2015). While contributing to the site’s early success, the ability to remain mostly anonymous and the opening of user-made subreddits invited the development of a series of high and low-profile subreddits that are (or have been) overtly and covertly violent (e.g./r/watchpeopledie), racist (e.g./r/coonland), anti-LGBTQ (e.g./r/DecencyAgainstLGBT), xenophobic (e.g./r/jewhate), or otherwise contentious (e.g./r/fatpeoplehate).<sup>2</sup> The introduction of stronger anti-harassment rules under Pao’s leadership in 2015, and subsequent updates to that policy,<sup>3</sup> marked a switch to a more active approach curtailing the user-generated communities. Reddit does not provide a list of banned subreddits, so the number of banned communities at any given time is unknown, but a subreddit dedicated to tracking the changing status of sanctioned communities on the site listed more than 200,000 subreddits as banned in March of 2020 (u/Elvis\_Interstellar, 2020).

While the large number of banned communities can be attributed to the ease of subreddit creation, Reddit’s willingness to police these spaces through community quarantines and bans is perhaps driven by a new and increasing reliance on external investors. Reddit is a private corporation, and through acquisition by Condé Nast and subsequent transition into an independent subsidiary of Advance Publications, Reddit has become, and remains, beholden to those who have invested roughly 550 USD million through four rounds of valuation (Saxena, 2019). With a most recent valuation of 3 USD billion (Saxena, 2019) and a rumored public offering (Rapier, 2017), Reddit has more aggressively enacted changes to user-content policies and increased community policing.

## Legal parameters

As a U.S. based social media organization, Reddit is not liable for content posted by users on its site. Regardless of the harm caused by users’ expression, whether it incites violence, harasses, or threatens another person, Reddit is not responsible. With the exception of content related to sex trafficking and copyright violations, Section 230 of the Communications Decency Act (CDA) says that providers of interactive computer services in the United States, such as ISPs or social media companies, shall not be treated as publishers and, therefore, are not responsible for what third parties do on their sites (*Communications Decency Act 1996*). Section 230 also says that if an intermediary does choose to police what its users say or do, it does not lose its safe harbor protections by doing so. In other words, choosing to delete some content does not automatically turn an intermediary into a publisher for legal purposes.

In Europe, however, countries have begun creating laws that hold social media companies such as Reddit accountable for the illegal content posted to their sites. In 2017, Germany passed the Network Enforcement Law, or NetzDG, which requires social media companies with more than two million users to remove or block access to reported content that violates restrictions against hate speech included in the German Criminal Code. Companies must remove “obvious hate speech” within

---

<sup>2</sup>All of the examples provided here are of banned subreddit communities.

<sup>3</sup>At the time of this writing, Reddit has announced that it is working on new policy changes reflective of widespread calls for reforms in the wake of protests brought about by the death of George Floyd and many other people of color in the United States. This announcement came with a slate of new subreddit bans, including the popular but very controversial/r/the\_donald.

24 hours of receiving a notification or risk a 50 USD million fine (Act to Improve Enforcement of the Law in the Social Networks, 2017). Under this framework, Reddit is required not just to quarantine but to completely remove content that meets the German definition of hate speech.

In addition to creating regulation to address harmful content on social media, many governments are working in partnership with social media organizations to address the issue. After the 2019 attack on two Mosques in New Zealand, global leaders met with executives from Facebook, Google, Twitter and other companies to compile a set of guidelines called the “Christchurch Call,” which sought to enact measures against extreme, violent and hateful rhetoric online. Notably, the United States did not sign the pledge. In the United States, Facebook, Google, Microsoft, and Twitter have partnered with the Anti-Defamation League (ADL) to create a Cyberhate Problem-Solving Lab to address the growing tide of online hate.

Despite these promising partnerships, Reddit is still free to regulate content in the United States in any way it sees fit. However, it is also not required to protect user expression. As a private virtual space, Reddit is not bound by the First Amendment of the United States Constitution. The terms of service users agree to for access to platforms such as Reddit serve as a private contract between the site and the user (Klonick, 2018). It is up to each individual platform to determine whether and how they remove or block access to content on their sites, a process called content moderation.

### **The process of content moderation**

Content moderation is best defined as a series of practices with shared characteristics, including rule-setting and the use of software and human moderators to regulate content, that are used to screen user-generated content to determine what will make it onto, or remain on, a social media platform (Gerrard, 2018; Roberts, 2019). Social media organizations regulate all kinds of potentially harmful content, including hate speech, violence or incitement, nudity (child and adult), sexual exploitation, sexual solicitation, suicide/self-injury, bullying, harassment, privacy violations, image privacy rights, promoting crime, or selling regulated goods.

This regulation process often includes three distinct elements: rule-making, automatic detection, and community flagging (Gillespie, 2018). Rule-making, which is also sometimes referred to as editorial review, is the first phase in the process of content moderation. In practice, this involves the creation and dissemination of community standards by social media platforms. Some sites, such as Instagram, choose to have community standards that prohibit posts that include nudity or racial slurs on their platforms. Others, such as Twitter, may ban threats of violence but not racial or ethnic slurs. Essentially, the rule-making phase of content moderation establishes the guidelines for what is considered acceptable and unacceptable content and behavior on each individual platform. To create an account and access a particular social media site, users must agree to the terms of services agreement, which is a contract between the user and the social media organization in which the user agrees to follow the community standards established by the organization. Failure to adhere to these rules can result in content removal, account suspension, or account termination, and detection of these violations is done either algorithmically (through automatic detection) or by other users (community flagging).

The second phase of the content moderation process is automatic detection, which refers to social media companies’ use of algorithms and artificial intelligence to remove content that violates the company’s community standards. Content can be removed before or after it has been posted. Sophisticated software is responsible for removing millions of pieces of content from various platforms. For example, in its 2019 Q1 Transparency Report, Facebook reported that it proactively removed 65% of hate speech on the site before users reported it (Facebook, 2019). While this software can be incredibly effective and is advocated for by some academics in the application of quarantine (Ullmann & Tomalin, 2020), it is not entirely neutral. Recent research suggests that there is a risk of racial bias in algorithms that detect hate speech or manage search functions (Noble, 2018; Sap, Card, Gabriel, Choi, & Smith, 2019). For example, an algorithm designed to detect hate speech may identify

and remove racial slurs but not gendered slurs because the software engineers that coded the algorithm do not consider terms such as “bitch” to be hate speech. When used to determine which accounts should be sanctioned or which content should be removed, algorithms may inadvertently identify innocuous content as problematic. Facebook and other social media platforms have run into this problem around posts that feature women breastfeeding. The algorithm marks this as nudity and removes it from the site, despite the fact that it is not sexual in nature, nor is it prohibited under the platform’s community guidelines. In addition, some anti-racist content has also been inadvertently removed by social media companies as hate speech, further demonstrating the potential bias baked into algorithms (Gynn, 2019).

The third and most visible component of content moderation is community flagging, which asks users to report content that they think violates a platform’s community standards. Generally, that reported content is then manually reviewed by employees or contractors of the social media organization to determine whether the reported content violates the rules. Some social media organizations, such as Facebook, outsource the review of content to other organizations/contractors who review the flagged content and make removal decisions. Workers in these roles are often dispersed globally at a variety of worksites, the work takes place in secret, workers are low-status, and paid very low wages (Roberts, 2016, 2019). They regularly suffer panic attacks and other mental health issues as a result of the hundreds of violent, hateful, or otherwise troubling posts they review each week. When a content reviewer determines that community standards have been violated, one (or more) of eight possible outcomes is usually the result: account restrictions, sending users warnings, content removal, disabling an account, deleting an account, working with law enforcement, removal from search, and removal from interface with a third party (Pater, Kim, Mynatt, & Fiesler, 2017).

### **Content moderation on Reddit.com**

From very near its creation, Reddit has had a two-tier content moderation system in place, which relies on both community moderators and site administrators (Massanari, 2015). The majority of content moderation on the site is done by subreddit moderators who, as volunteers and active members of the communities they moderate, establish the rules for the individual subreddits they moderate and then monitor the content posted to enforce those rules (Fiesler, Jiang, McCann, Frye, & Brubaker, 2017). Moderators have the power to remove/delete posts, apply sanctions to users within a community, ban users from a community, employ algorithmically-managed content controls, and provide escalated reporting to Reddit about user behavior (Massanari, 2017). In this way, Reddit essentially outsources all three elements of the content moderation process (rule-making, automatic detection, and community flagging) to subreddit community moderators.

When content violates the site’s harassment policy which prohibits “unwelcome content,” including involuntary pornography, personal or confidential information, encouraging or inciting violence, or content that “threatens, harasses, or bullies or encourages others to do so,” the company may step in to override an individual community’s self-policing (Reddit Inc., 2020a; Reddit Inc., 2020b)). In the past, user complaints and/or public pressure have motivated Reddit to either ban a subreddit entirely, as it recently did with *r/the\_donald*, or to quarantine the subreddit (Allyn, 2020). Quarantine prevents posts from appearing on the home page, *r/all* and adds a warning page featuring boilerplate language about the potentially offensive content featured on the page, which users must click-through in order to access the quarantined site. Unlike many other social media platforms, Reddit does not (yet) employ artificial intelligence or software content moderators to scour the entire site for violations to its community standards.

The subreddit *r/TheRedPill* was quarantined in September of 2018 for being “dedicated to shocking or highly offensive content” (Controversial Reddit communities, 2020). Unlike banning the subreddit entirely, this approach allows Reddit to keep participants in contentious communities active on the site, which drives advertising and, ultimately, revenue, while appearing to publicly denounce their existence and impact. Our analysis of current posts on *r/TheRedPill* (2019) indicated that

despite the quarantine, community members on the subreddit continue to violate the existing harassment policy by calling for domination and control over the social, physical, and sexual spaces of women.

## The ethics of content moderation

Social media companies are currently developing ethical guidelines for content moderation; however, as of yet, there is not a single agreed-upon code that all platforms follow. Legally, each corporation is free to determine whether and how it will approach the process of content moderation. Recognizing the issues that can arise from both a lack of action or too much censorship, scholars have begun to offer ethical frameworks to guide social media organizations' in their content moderation efforts.

In his 2017 article, "Speech, Harm, and the Duties of Digital Intermediaries: Conceptualizing Platform Ethics," scholar Brett Johnson explores the question of whether social media platforms have a duty to prevent harm caused by vile speech on their platforms, or whether their commitment should be to free expression, and thus should work to ensure as much speech as possible is published on their platforms. To answer this question, Johnson (2017) offers two frameworks. One is grounded in a duty to promote speech, and another is rooted in digital intermediaries' duty to prevent harm to users whose data and content these companies commodify in order to make a profit.

The latter approach argues that respecting the dignity of users is done through preventing the harm they might endure on a particular platform. Unlike traditional media, digital intermediaries are in a unique position to mitigate harm through content removal. Failure on their part to act when necessary subjects their users to unnecessary suffering and in doing so disrespects their dignity (Johnson, 2017). To respect the dignity of their users, digital intermediaries should act as active custodians of the public discourse that takes place on their platforms. They should "continue to respond to users' requests to review and remove harmful content but they also should employ both the people and the technical capabilities to seek out and eradicate examples of harmful user-generated content before user complaints arise" (Johnson, 2017, p. 22–23).

Another framework for ethical conduct in social media content moderation are the Santa Clara Principles. Developed during two Content Moderation at Scale conferences in 2018 by a diverse set of stakeholder groups and experts (including the ACLU, the Center for Democracy & Technology, the Electronic Frontier Foundation, New America's Open Technology Institute, and a number of academic experts), they outline "minimum levels of transparency and accountability," and propose "three principles as initial steps that companies engaged in content moderation should take to provide meaningful due process to impacted speakers and better ensure that the enforcement of their content guidelines is fair, unbiased, proportional, and respectful of users' rights" ([santaclaraprinciples.org](http://santaclaraprinciples.org), 2020). The first principle, numbers, states that companies should publish data about how many posts were removed and how many accounts sanctioned through their policing of content guidelines. This data should be published regularly and be openly licensed. The second principle, notice, compels intermediaries to provide notice to any user affected by content moderation and this, along with detailed information about content guidelines, should be open and easily accessible to users. The third and final principle would grant users who have been affected by content moderation the right to an appeal that is transparent and allows the user to defend their content in a meaningful way.

The notion of transparency runs throughout all three of the directives that make up the Santa Clara Principles. Transparency is rooted in the work of Immanuel Kant, who linked the integrity of action to human dignity (Plaisance, 2007). By honoring truth and honesty through transparent behavior, we fulfill our obligation to respect the rational agency and free will of everyone we communicate with. Transparency has long been a pillar of media ethics and so too should be relevant to any discussion of social media ethics. Transparent behavior "presumes an openness in communication and serves a reasonable expectation of forthright exchange when parties have a legitimate stake in the possible outcomes or effects of the communicative act" (Plaisance, 2007, p. 188). Thus, being transparent is one

of the ways social media organizations can respect the dignity of their users. Therefore, it is essential for social media companies to be transparent about their rules, the application of those rules, and the impact the process of moderation has on users' free expression.

In addition to proposing formal frameworks such as the Santa Clara Principles to guide content moderators, scholars have begun to explore how traditional ethical frameworks can apply to social media content management decisions. Shannon Bowen (2013) argues that Kantian deontology or duty-based ethics should be used to guide public relations professionals' social media interactions with their organization's stakeholders. Bowen's model asks professional communicators to uphold universal moral principles, as well as the dignity and respect of posters and their audience by communicating with good intention. Her ethical guidelines for using social media ask public relations practitioners to rely on universal principles such as dignity, fairness, honesty, transparency, and respect when communicating via social media (Bowen, 2013, p. 20)

Looking at various case studies, Bowen describes how certain organizations have failed to do this by using social media to deceive stakeholders. Although she is looking at product endorsements or promotional accounts rather than content moderation, the notion of applying duty-based ethics to social media is a valuable one. If Reddit content moderators were being directed to respect the dignity of posters and audiences, it's likely that they would be encouraged to remove *r/TheRedPill* because it degrades women and minimizes their dignity. Removing *r/TheRedPill* entirely, rather than simply quarantining it, would go further toward upholding Bowen's universal ethical principles of dignity, respect, and fairness.

Another useful tool for considering questions about content removal is John Rawls (1971) "veil of ignorance" device. Here, decision-makers are asked to ignore their own personal circumstances, including their social or economic position, when creating a social contract to afford all parties equal consideration (Rawls, 1971). Two principles are central to this framework. First is the liberty principle, which seeks maximum liberty for all, while avoiding intrusion upon the freedom of others. The second is the difference principle, which seeks to give all people an equal chance to prosper, despite social and economic differences. By setting aside one's own position in this process, people will be more likely to develop rules and regulations that are fair and just.

In the case of social media, adopting this mentality would call for companies to set aside their own position, in which the economic benefits and drawbacks of a content removal are paramount, and instead consider the impact on people. Adopting the liberty principle would guide content moderators to favor freedoms only up until the point that they infringed on others' ability to participate on the site in meaningful ways.

This framework seems particularly useful in light of concerns about the financial benefits that are potentially tied to allowing hateful content to remain on these platforms. YouTube and Facebook, for example, have both been accused of allowing hateful content to remain on their site as a way of attracting viewers and users, which can be translated into advertising dollars (Dwoskin, 2019).

Given the fact that social media organizations are publicly traded companies operating worldwide, it also makes sense for them to consider their commitment to corporate social responsibility as they navigate the ethical questions raised by their content moderation practices. Corporate social responsibility refers to the relationships, obligations, and duties that exist between a corporation and its stakeholders and that guide business behavior (Steiner, 1972). To be considered a good corporate citizen, companies must fulfill their economic, legal, ethical, and philanthropic responsibilities (Carroll, 2016). These four elements make up Carroll's pyramid of corporate social responsibility. Each component in the pyramid addresses different stakeholders in terms of the varying priorities in which the stakeholders might be affected. According to Carroll, the pyramid is intended to be seen as a dynamic, adaptable framework that constantly considers and reconsiders both stakeholders and sustainability (Carroll, 2016). Oftentimes, the commitments made to stakeholders are evident in the company's mission and values. In the case of Reddit, the company's mission statement, according to a 2017 article in *Variety*, is "On Reddit, users can be themselves, learn about the world around them, and be entertained by the content created and shared by our global community" (Spangler, 2017).



Notably, the company's mission statement is not listed on the corporate site. To fulfill this duty and their commitment to stakeholders, Reddit must enact its promise to allow users to be themselves. From this perspective, the company should allow subreddits such as *r/TheRedPill* to remain on the site in order to fulfill its mission.

Clearly, there are many ways to think about the ethics of content moderation. Each of the approaches outlined here offers valuable insights that can and should be used to guide social media organizations such as Reddit as they wrestle with difficult decisions regarding whether to remove hateful or harassing content. From this review of various scholars' work, several key tenets have emerged, which can provide a framework for addressing the many ethical questions that surround the process of social media content regulation. The first is transparency. Social media organizations can respect the dignity of users by being transparent in the publication of their rules and the reporting of their efforts at regulation. Bowen's work offers ethical guidelines for social media interactions that focus on our duty to uphold certain moral principles, such as respect, honesty, and fairness. Rawls' veil of ignorance theory is another effective tool social media content moderators could adopt to better consider all of the parties impacted by a particular decision. By taking away their own positionality, this approach would arguably move social media content moderators toward a decision-making framework that benefits the stakeholders with the least power. Finally, along those lines, a corporate social responsibility model would call on social media to consider the impact of their actions on all stakeholders, including shareholders, advertisers, users, and society. Taking each of these approaches into account, we offer a set of guidelines that content restriction decisions should abide by:

- (1) Social media organizations should respect the dignity of their users.
- (2) Social media organizations should act as active custodians for democratic discourse on their platforms.
- (3) Decisions regarding content removal should be transparent.
- (4) Companies should establish a set of moral principles they seek to uphold with their decisions. These may include respect, transparency, honesty, and fairness.
- (5) Social media organizations should divorce themselves, as much as possible, from only considering what's best for the company and instead approach decision-making from as neutral of a position as possible.
- (6) Social media organizations should consider the impact of their content removal decisions on all stakeholder groups, particularly those that have little power or have historically been marginalized.

By drawing these guidelines from existing research, we hope to provide a framework that can be utilized to determine whether social media companies such as Reddit should remove, quarantine, or simply ignore subreddits such as *r/TheRedPill* that prominently feature hateful and harassing content.

### **Is Reddit's decision to quarantine *r/TheRedPill* ethical?**

In this section, we will apply the guidelines proposed here to Reddit's decision to quarantine the Red Pill subreddit, to determine whether or not this approach should be considered an ethical course of action, whether it would be better to remove the site all together or to leave it untouched.

The first principle proposed draws on Johnson's (2017) work and encourages us to respect the dignity of users. While it is possible to interpret this solely as respecting the dignity of *r/TheRedPill* community members, we read this as a directive to respect the dignity of ALL reddit users, including the millions of women who use the site daily. Quarantining *r/TheRedPill* subreddit removes this content from the main *r/all* feed so that those targeted or offended by the expression of *r/TheRedPill* community members are not inadvertently exposed to it. However, this leads to the question of whether allowing it to stay on the site at all, even behind a security screen, is respectful of all users'

dignity. We would argue that it is not. The content on/r/TheRedPill subjugates women and can serve to either silence them (Citron & Norton, 2011), or make discrimination against them more palatable. For this reason, we believe that respecting users' dignity means removing the subreddit entirely.

The next element in the decision-making model instructs social media companies to act as custodians for democratic discourse on their site. Digital intermediaries such as Reddit are vital in facilitating democratic discourse, which is at its best when more speech by individuals is allowed rather than restricted (Johnson, 2017). Following this line of thinking, it would be best for Reddit to leave/r/TheRedPill untouched, rather than quarantining it or removing it. Democracy requires an open exchange of ideas. Silencing objectionable speech before it sees the light of day prevents it from reaching the marketplace of ideas, where it can be tested against other perspectives and where hopefully, the truth will ultimately emerge. Moreover, to govern ourselves effectively, we need access to all information, including that which we find offensive. In order to promote discourse and facilitate the democratic process, the/r/TheRedPill subreddit should not be removed from the site and should not be quarantined.

Transparency is the third tenet of the proposed framework for ethical decision-making in social media content moderation. Here the focus is less on what Reddit decides and more on how it communicates those decisions. In this sense, the company is free to quarantine, remove, or leave the/r/TheRedPill provided that it makes that information readily available to users. Transparency is an area that Reddit could greatly improve upon. While fans of the site maintain a Wikipedia page and subreddit that track removed and quarantined subreddits (Controversial Reddit communities, 2020), this information is not available on Reddit's corporate home page. Redditinc.com contains information for potential advertisers and employees, and even features a "Reddit by the numbers" section that details the number of active communities or daily users but provides no information about the company's content moderation actions. This would be a perfect location for Reddit to include a detailed transparency report that outlines which subreddits have been removed and why, and which subreddits are under quarantine and why. Being transparent about these decisions is not only a far more ethical approach than the secrecy currently being employed, but it is also likely to be beneficial for advertisers and potential investors to have access to this information.

The fourth portion of the proposed model asks social media organizations to identify a set of guiding moral principles that can help facilitate decisions regarding content removal. Reddit Inc.'s website does not list a company mission or values in their "about" section. However, according to a 2015 blog post, Reddit Inc.'s core values are, or at least were: (1) To remember the human by being authentic, passionate, and empathetic. Treat others as you would in person. Champion diversity. Default to transparency and be honest. (2) Give people voices by creating a safe space to encourage participation. Allow freedom of expression. Be stewards not dictators and let communities own themselves. (3) Respect anonymity and privacy (4) Embrace experimentation (5) Make deliberate decisions (6) Be doers (7) The spirit of the Lambeosaurus (dinosaur) embiggins us all. Work is better when you're having fun. Don't take things too seriously (Reddit Inc., 2015).

Admittedly there are a lot of potentially conflicting ideas included in these value statements and their subpoints. If Reddit were to prioritize championing diversity, then it would likely remove or at least continue to quarantine/r/TheRedPill as much of the information posted in the community devalues women as unintelligent, sexually promiscuous, and essentially, sub-human. However, if the company chose to focus on its support for free expression, as a guiding principle, it would likely decide not to quarantine or remove/r/TheRedPill.

While the values listed in the 2015 blog post are a good start, Reddit would likely benefit from narrowing these principles to include only those that are most essential to the organization and its work and using those values to drive content moderation decisions.

The fifth element of our ethical decision-making model asks organizations to adopt a veil of ignorance and make decisions as if they were blind to their own social and economic circumstances. Translating this into a daily decision-making tool, Reddit's content moderators could assess various situations by asking themselves which outcomes allow all parties to prosper equally. If Reddit ignored

its own position, the company would likely decide to remove the/r/TheRedPill because it would see the hateful and harassing content as deleterious to women users of the site (estimated to be several million), rather than a way to maintain engagement from a mid-sized subreddit community (less than 400,000). By minimizing the decision-making power of the potential impact on advertising revenue, or a potential public offering, content moderators at Reddit could better consider the point of view of the people who are demeaned by the content posted on/r/TheRedPill subreddit.

The final tenet of the proposed ethical framework draws on the stakeholder model of corporate social responsibility (Carroll, 2016) to encourage social media organizations such as Reddit to consider the impact of its decisions on all stakeholder groups, particularly those whose identities have historically been marginalized. In the past, Reddit has struggled to adopt this perspective. Rather than proactively removing communities that feature racist, misogynistic, and/or homophobic content, Reddit has waited until public pressure forced them to do so. For example, in response to the murder of George Floyd by Minneapolis police officers, and the nationwide BlackLivesMatter protests that followed, Reddit made a decision to remove the subreddits/r/The\_Donald and/r/ChapoTrapHouse, along with 2,000 other communities, after updating its content policy to more explicitly ban hate speech. Rather than wait until these issues enter the cultural zeitgeist, Reddit should constantly be thinking and re-thinking about how the content on its site impacts both its users and society more broadly. It stands to reason that if preventing violence against women suddenly becomes in vogue, Reddit would undoubtedly act quickly to remove/r/TheRedPill. Rather than wait, Reddit should immediately focus on how the content on its site impacts all stakeholders, paying particular attention to those whose identities have traditionally been marginalized. Following that line of reason, it is likely Reddit would decide to remove/r/TheRedPill from its site.

## Conclusion

The application of the proposed decision-making model to/r/TheRedPill indicates that the most ethical course of action is to remove the community from the site. This would respect users' dignity and ultimately minimize the negative consequences this kind of content has on members of traditionally marginalized communities. However, there are reasonable arguments in favor of leaving the site up and under quarantine, such as promoting democratic discourse and supporting freedom of expression. As with most decisions, there is no easy answer. Instead, we hope that this framework provides social media organizations with guideposts for addressing complex questions about content removal. To be successful, the process of content moderation must be nuanced and balanced, able to simultaneously account for the dignity of users and the right of those users to freely express themselves. Fortunately, social media platforms such as Reddit are free to engage in this process in whatever way makes the most sense to them. We encourage Reddit and other social media companies to take seriously the role they have assumed as arbiters of public discourse and continue to work toward achieving the nuance and balance its task demands. Social media organizations' decisions reverberate throughout society, and therefore, they must consider the ethical foundations that drive their content moderation processes.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## ORCID

Caitlin Ring Carlson  <http://orcid.org/0000-0002-3286-2436>  
Luc S. Cousineau  <http://orcid.org/0000-0002-6475-8852>

## References

- Act to Improve Enforcement of the Law in the Social Networks. (2017). Retrieved from [https://www.gesetze-im-internet.de/englisch\\_stgb/englisch\\_stgb.html](https://www.gesetze-im-internet.de/englisch_stgb/englisch_stgb.html)
- Allyn, B. (2020, June 29). Reddit bans the\_donald, forum of nearly 800,000 trump fans, over abusive posts. *NPR*. Retrieved from [https://www.npr.org/2020/06/29/884819923/reddit-bans-the\\_donald-forum-of-nearly-800-000-trump-fans-over-abusive-posts](https://www.npr.org/2020/06/29/884819923/reddit-bans-the_donald-forum-of-nearly-800-000-trump-fans-over-abusive-posts)
- Bowen, S. A. (2013). Using classic social media cases to distill ethical guidelines for digital engagement. *Journal of Mass Media Ethics*, 28(2), 119–133. doi:10.1080/08900523.2013.793523
- Carroll, A. B. (2016). Carroll's pyramid of CSR: Taking another look. *International Journal of Corporate Social Responsibility*, 1, 3. doi:10.1186/s40991-016-0004-6
- Citron, D. K., & Norton, H. (2011). Intermediaries and hate speech: Fostering digital citizenship for our information age. *Boston University Law Review*, 91, 1435–1484.
- Communications Decency Act 1996, 47 U.S. § 230(c)(2).
- Controversial Reddit communities. (2020). *Wikipedia*. Retrieved from [https://en.wikipedia.org/w/index.php?title=Controversial\\_Reddit\\_communities&oldid=959288393](https://en.wikipedia.org/w/index.php?title=Controversial_Reddit_communities&oldid=959288393)
- Dvoskin, E. (2019, August 9). YouTube's arbitrary standards: Stars keep making money even after breaking the rules. *The Washington Post*. Retrieved from <https://www.washingtonpost.com/technology/2019/08/09/youtubes-arbitrary-standards-stars-keep-making-money-even-after-breaking-rules/>
- Elvis Interstellar. (2020). Updated list of all known banned sub reddits sorted by reason and alphabetically. reclassified. Retrieved Sept. 7, 2020, from Reddit [https://www.reddit.com/r/reclassified/comments/fg3608/updated\\_list\\_of\\_all\\_known\\_banned\\_subreddits/](https://www.reddit.com/r/reclassified/comments/fg3608/updated_list_of_all_known_banned_subreddits/)
- Facebook. (2019). *Transparency report*. Retrieved from <https://transparency.facebook.com/community-standards-enforcement#hate-speech>
- Fiesler, C., Jiang, J., McCann, J., Frye, K., & Brubaker, J. (2017). Reddit rules! Characterizing an ecosystem of governance. *Proceedings of the association for the advancement of artificial intelligence*, San Francisco, California.
- Futrelle, D. (2011, April 29). WTF is a MGTOW? A Glossary. *We Hunted the Mammoth*. Retrieved from <http://www.wehuntedthemammoth.com/wtf-is-a-mgtow-a-glossary/>
- Gerrard, Y. (2018). Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society*, 20(12), 4492–4511. doi:10.1177/1461444818776611
- Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. New Haven and London: Yale University Press.
- Ging, D. (2019). Alphas, betas, and incels: Theorizing the masculinities of the manosphere. *Men and Masculinities*, 22(4), 638–657. doi:10.1177/1097184X17706401
- Gotell, L., & Dutton, E. (2016). Sexual violence in the 'manosphere': Antifeminist men's rights discourses on rape. *International Journal for Crime, Justice and Social Democracy*, 5(2), 65. doi:10.5204/ijcsd.v5i2.310
- Guynn, J. (2019, August 24). Facebook while black: Users call it getting 'Zucked,' say talking about racism is censored as hate speech. *USA Today*. Retrieved from <https://www.usatoday.com/story/news/2019/04/24/facebook-while-black-zucked-users-say-they-get-blocked-racism-discussion/2859593002/>
- Johnson, B. G. (2017). Speech, harm, and the duties of digital intermediaries: Conceptualizing platform ethics. *Journal of Mass Media Ethics*, 32(1), 16–27. doi:10.1080/23736992.2016.1258991
- Klonick, K. (2018). The new governors: The people, rules, and processes governing online speech. *Harvard Law Review*, 131, 1598–1690.
- Marwick, A. E. (2017). Scandal or sex crime? Gendered privacy and the celebrity nude photo leaks. *Ethics and Information Technology*, 19(3), 177–191.
- Massanari, A. (2017). #Gamergate and The Fapping: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, 19(3), 329–346. doi:10.1177/1461444815608807
- Massanari, A. L. (2015). *Participatory culture, community, and play: Learning from Reddit*. New York, NY: Peter Lang.
- Moreno, J., Pao, E., & Ohanian, A. (2015, May 14). *Promote ideas, protect people*. Upvoted. Retrieved from <https://redditblog.com/2015/05/14/promote-ideas-protect-people/>
- Noble, S. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York, NY: NYU Press.
- Ohanian, A. (2016). *Without their permission: The story of Reddit and a blueprint for how to change the world*. New York, NY: Grand Central Publishing.
- Pater, J. A., Kim, M. K., Mynatt, E. D., & Fiesler, C. (2017). Characterizations of online harassment: Comparing policies across social media platforms. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 1501–1513).
- Plaisance, P. L. (2007). Transparency: An assessment of the Kantian roots of a key element in media ethics practice. *Journal of Mass Media Ethics*, 22(2–3), 187–207.
- Rapier, G. (2017, November 14). Reddit is reportedly considering an IPO. *Business Insider*. Retrieved from <https://www.businessinsider.com/reddit-is-reportedly-considering-an-ipo-2017-11?op=1>
- Rawls, J. (1971). *A theory of justice*. Oxford: Oxford University Press.

- /r/TheRedPill. (2019, October 3). R/TheRedPill. Retrieved October 3, 2019, from Reddit website: <https://www.reddit.com/r/TheRedPill/>
- Reddit Inc. (2015, May 5). *We're sharing our company's core values with the world*. Retrieved from <https://redditblog.com/2015/05/06/were-sharing-our-companys-core-values-with-the-world/>
- Reddit Inc. (2020a, June 1). *Do not threaten, harass, or bully*. Reddit Help. Retrieved from <https://www.reddithelp.com/en/categories/rules-reporting/account-and-community-restrictions/do-not-threaten-harass-or-bully>
- Reddit Inc. (2020b, April 10). *Content policy*. Retrieved from <https://www.redditinc.com/policies/content-policy>
- Reddit Metrics. (2020, July 21). New subreddits by month—Reddit history. Retrieved July 21, 2020, from <https://redditmetrics.com/history/month>
- Reeve, E. (2018, August 2). This is what the life of an incel looks like. *Vice News*. Retrieved from [https://news.vice.com/en\\_us/article/7xqw3g/this-is-what-the-life-of-an-incel-looks-like](https://news.vice.com/en_us/article/7xqw3g/this-is-what-the-life-of-an-incel-looks-like)
- Roberts, S. T. (2016). Commercial content moderation: Digital laborers' dirty work. In S. U. Noble & B. M. Tynes (Eds.), *The intersectional internet: Race, sex, class and culture online* pp. (147–160). New York, NY: Peter Lang Publishing.
- Roberts, S. T. (2019). *Behind the screen: Content moderation in the shadows of social media*. New Haven, CT: Yale University Press.
- santaclaraprinciples.org. (2020). *Santa clara principles on transparency and accountability in content moderation*. Santa Clara Principles. Retrieved from <https://santaclaraprinciples.org>
- Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). *Proceedings from the 57th annual meeting of the association for computational linguistics: The risk of racial bias in hate speech detection* (pp. 1668–1678).
- Saxena, A. (2019, February 11). Reddit valued at \$3 billion after raising \$300 million in latest funding round. Reuters. Retrieved from <https://www.reuters.com/article/us-reddit-funding-idUSKCN1Q020W>
- Spangler, T. (2017, August 1). Reddit has \$1.8 billion valuation after chat-room site banks \$200 million in funding. *Variety*. Retrieved from <https://variety.com/2017/digital/news/reddit-1-8-billion-valuation-funding-1202512082/>
- Steiner, G. A. (1972). Social policies for business. *California Management Review*, 15(2), 17–24.
- Ullmann, S., & Tomalin, M. (2020). Quarantining online hate speech: Technical and ethical perspectives. *Ethics and Information Technology*, 22, 69–80. doi:10.1007/s10676-019-09516-z