The Human Right to Freedom of Expression in the age of social media

Aimée M.R. Povel

supervised by Alejandra Mancilla

Thesis presented for the degree of Master in Philosophy
Department of Philosophy, Classics, History of Art and Ideas
Faculty of Humanities

UNIVERSITY OF OSLO

June 2020

Acknowledgements

I would like to thank Alejandra Mancilla, Associate Professor at the University of Oslo, for inspiring supervision over the period of January 2019 until June 2020. Thank you so much for true support, for teaching me about the subject matter and the writing process, and for challenging me in the right way. I hope the thesis does your guidance justice.

I would like to thank my dear friend Karin Mæland, for inspiring me as a philosopher, for never tiring of talking about the topic of my thesis with me, for proof-reading, and for unwavering friendship and support.

I would like to thank all my professors at IFIKK, Arne Johan Vetlesen, Øyvind Rabbås, Panos Dimas, Herman Wright Cappelen, Franco Trivigno, for all they teach and share, and studies administrator Caroline Hansen for her admirable capacity and friendliness.

I would like to thank friends, family, and colleagues for warm support and for tolerating what basically turned into two years of me being absent.

Abstract

Should we rethink the Human Right to Freedom of Expression [HRFE] in the age of social media? In this thesis I have reevaluated the philosophical foundation of the HRFE in how it is applicable to Facebook. I argue that the traditional Millian account of freedom of expression [FE] can ground a contemporary HRFE integrating all forms of expression, media, and public discourse in contemporary liberal democratic society, including social media. The problem is that social media like Facebook on some essential points do not meet the traditional conditions for FE. I argue that large platforms for expression like social media should meet these conditions. I then hint to some ways in which we could seek to harmonize social media with the traditional account of FE.

The philosophical contribution this thesis purports to make is to create awareness that current "online" expression is divorced from our traditional "offline" conception of FE and that this harms liberal democratic legitimacy.

The way forward I see to establishing a philosophical account of an integrated HRFE, is by taking as a point of departure that it should be grounded in offline society and its national identities. Then, issues concerning identity, conceptions of harm, and how to account for globalized online communication can be addressed as extensions and adaptations of the traditional account, adjusted to contemporary society. To achieve this, society should stop the self-regulation of tech companies and allow national governments some regulation of online expression to include it in the protection of an integrated HRFE.

Index

0 Introduction 1
1 The Human Right to Freedom of Expression in the law 4
1.1 The form of the right 4
1.2 The legal content and scope of the HRFE 6
2 The HRFE as a protection of liberal democracy 10
2.1 A basic conception of liberal democracy 10
2.2 The traditional liberal defense of freedom of expression 11
2.3 Five conditions for freedom of expression 28
3 How Facebook makes the traditional account of FE run into trouble 30
3.1 The expression is not harmful 31
3.2 One person, one voice 44
3.3 The openness to dialogue of honest speakers 47
3.4 Real and morally accountable people 54
3.5 A closed, physical society 61
3.6 Summary 68
4 Should expression on social media be regulated by liberal democracies? 70
4.1 Some ideas on regulation of social media 72
5 Conclusion 77
6 Literature 80

0 Introduction

During the hearing of Mark Zuckerberg, CEO of Facebook Inc., in the American senate in 2018, senator Graham asked Zuckerberg, “what do I tell my constituents about why we should let you self-regulate?” (CBSN, 2018). The context for the question was Facebook's (hereinafter, FB) unlawful sharing of FB user information with third parties, a breach of privacy law for which FB was penalized with the largest fine in history (the Washington Post, 2019). FB is a social media platform, facilitating a global online network of users. Zuckerberg replied that FB is open for regulation if it is the right kind of regulation, without specifying what right regulation could look like (CBSN, 2018).

Why should we let social media companies like FB self-regulate policies and practice regarding (freedom of) expression online? And if we don't accept self-regulation, what should the right regulation address and look like? The answer to these questions affects the Human Right to Freedom of Expression of over two billion FB users worldwide, every day, directly. But the answer also affects society offline, as FB's policies facilitate harm caused by third parties, targeted voter manipulation in national elections, and has resulted in real physical harm to individuals in the real world (Amer & Noujaim, 2019).

In this thesis I discuss FB as a representative case of how tech companies that create and manage social media platforms affect the role of the Human Right to Freedom of Expression (hereinafter, HRFE) in protecting liberal democracy.

It seems we lack a philosophical account of the role social media play in the contemporary practice of the HRFE. Our conception of freedom of expression (hereinafter, FE) is founded upon premises concerning what liberal democratic society looks like, that duties correlating with a the HRFE (mostly) belong to the state, and that the HRFE has an essential role in protecting liberal democratic society. In this thesis I will evaluate how social media challenge these premises and analyze the discrepancy between the traditional account of FE and expression on social media. The question my thesis focuses on is how the HRFE and its role in protecting liberal democracy should be realized in the face of social media.

In chapter one, I give an account of the legal conception of the HRFE as a claim right with correlated duties.

To evaluate my question concerning the HRFE in the age of social media, I use a normative account of freedom of expression based on three philosophers in the liberal tradition, namely John Stuart Mill, Ronald Dworkin, and Jeremy Waldron.

In chapter two, I start with John Stuart Mill's views as presented in *On Liberty*, first published in 1859 (Mill, 1989). Mill's view is that civil liberty and human progress require the greatest possible scope of FE, restricted only to protect others from harm (1989, p.13 and p.15). This is called the harm principle. For a discussion of legitimate restrictions of FE based on Mill's harm principle, I focus on arguments given by Ronald Dworkin and Jeremy Waldron.

Dworkin argues for a near-absolute scope of FE, delegating the protection from harm to laws that do not affect FE (Dworkin, 2006, p.132). Dworkin claims that FE should be an inviolable principle, since it is essential for legitimizing the outcomes of liberal democracy. Restricting FE would violate the principle that gives it its power (Dworkin, 2009, p.ix).

According to Dworkin, the liberal aspect of liberal democracy should trump other concerns. For Waldron, it is not the liberal, but the egalitarian values liberal democracy represent that should inform FE. According to Waldron, then, the equal status of dignity of citizens requires that minorities and vulnerable groups should be protected from the harm in hate speech (Waldron, 2012, pp.4-5).

Mill, Dworkin, and Waldron agree that FE is essential as a protection of liberal democracy. This is the role of the traditional account of FE I discuss in my evaluation of social media like FB and their treatment of the HRFE. But these three philosophers also share that the account they provide of FE has "offline" society as its premise. "Offline" society is society before the World Wide Web became a defining factor. In this thesis I sometimes refer to contemporary society as "online society". The question is whether their traditional account of FE holds up in the face of social media.

In chapter three, I discuss FB as a social media platform representative of social media today. I show how FB works, and how each of the conditions of the traditional account of FE is challenged. I discuss these questions in relation to five central conditions for the traditional account of FE to be protected as a human right, namely: the expression is not harmful; one person, one voice; the openness to dialogue of honest speakers; real and morally accountable people; and a closed, physical society.

FB's individually tailored user experience is problematic when considered from the perspective of what FE entails and is meant to protect. FB's content moderation interferes with a user's free and undistorted access to the intended audience of one's expression, plus it restricts a symmetrical access to the real diversity of available expressions. From a Millian viewpoint, in doing this it violates both the right to FE and its function in relation to liberal democracy. I argue that social media like FB restrict expressions on its social media platform in a way that is not aligned with the traditional account of FE. Furthermore, FB's speech policies effectively

remove expressions from the democratic processes and functions they are meant to be part of and protect. Therefore, FB restricts their users' FE and negatively affects liberal democracy.

In chapter four I consider what the discrepancies between the traditional account of FE and expression on social media like FB should mean for the HRFE. If studying FB as a case shows us that online expression does not meet the conditions of FE, and assuming social media have come to stay, how can we accomplish the conditions for FE online? This thesis aims to highlight that the difference between what we intent with the HRFE and the reality of online expression is a problem area for the practice of the HRFE.

Should we conclude that the current situation provides us with enough goods to accept it as it is, letting the HRFE govern offline liberal democracy, and leaving online expression under the moderation of self-regulated tech companies without enforcement of the HRFE? Or should we rethink the HRFE? If not, should we impose the HRFE on social media through governmental regulation?

I believe we should regulate social media like FB into protecting their users' HRFE. This would mean no moderation of online expression other than to protect others from harm. The most likely way of achieving this is through active intervention of liberal democratic governments to regulate online media.

1 The Human Right to Freedom of Expression in the law

Freedom of expression is recognized as an indispensable fundamental right in liberal democracies. Conceptions of the right do differ among states, especially concerning what counts as a legitimate restriction of the scope of the right. To be able to refer to the international recognition of the right rather than the national differences, I will ground this thesis in the understanding of FE we internationally have in the form of the Human Right to Freedom of Expression.

In this chapter I discuss the HRFE's legal form, content and scope, and what the wording of the right is in the main legal documents. These legal documents are the Universal Declaration of Human Rights [UDHR] (1948) and the European Convention on Human Rights [ECHR] (1950).

Different legal documents use the phrase "freedom of expression", echoing the UDHR. The American First Amendment refers to FE as "freedom of speech" in its Bill of Rights (U.S. Const., amend. I). Literature discussing FE also uses "free speech (principle)", "freedom of information", and other terms seemingly interchangeably. In this thesis I will use "freedom of expression" or "free speech" interchangeably when referring to FE as a philosophical principle. I will use "HRFE" when referring to the human right as enshrined in the law.

1.1 The form of the right

A central philosophical analytical tool for understanding the relation between rights (or claims) and duties is the Hohfeldian analysis (Hohfeld, 1913). Hohfeld bases his analytical model on the relations of opposites and correlations of rights and duties. The four Hohfeldian *incidents* that constitute rights are: claims, duties, liberties, and no-claims. To illustrate their correlations: if Jane has a claim against Joe, Joe has a duty towards Jane. If Jane is at liberty to perform an action, Joe has no claim on Jane not to perform this action. In this way, having a liberty or claim right for Jane is to be in relation with Joe regarding a duty to (non)action.[1]

[1] Shue agrees with the Hohfeldian focus on duties correlating with claim rights: «[I]t is essential to a right that it is a demand upon others, however difficult it is to specify exactly which others» (Shue, 1980, p.16). Another definition of claim rights that bases itself on the correlation with duties is given by Raz: "X has a [claim] right' if and only if… an aspect of X's well-being (his interest) is a sufficient reason for holding some other person(s) to be under a duty" (Raz, 1986, p.166). Claim rights not only give the right holder the right to an action (or inaction), but also the claim to the correlating duty.

Hohfeld recognizes that “rights”, even specified as four incidents, is a term that has a wide range of possible meanings. Therefore, he proposes to take the understanding of the correlative duty to narrow and clarify its scope. Hohfeld says that “if X has a right against Y that he shall stay off the former's land, the correlative (and equivalent) is that Y is under a duty toward X to stay off the place” (Hohfeld, 1913, p.32). So, the content and scope of a right can be specified through the claim it provides to the related duty of the other party.

Important for our conception of the HRFE is the understanding that it is not only a liberty, but a claim right. Thus, it correlates with a duty. Duties correlating with the HRFE can be, on Shue’s view of (basic) rights, defined into three aspects.[2] These are to *avoid* interference (noninterference), generally referred to as a negative duty, or to *protect* the HRFE against interference, which is an example of a positive duty, or to *assist* in having the HRFE exercised (Shue, 1980, pp.38-9).

An argument against the legitimacy of human rights concerns the feasibility of claims to correlating duties. Duties need to be realizable by duty-bearers. As Gilabert argues, “You cannot have a duty to do what you cannot do. Since rights imply obligations and obligations imply feasibility of compliance, infeasibility of compliance with certain obligations implies the absence of these obligations and the absence of their correlative rights” (Gilabert, 2009, p.659). So, if a human right implies a correlating duty, this duty must be feasible for the claim to be a right. Though at least for a state, the HRFE involves positive duties and costs, I don’t think the HRFE has problems with duty-bearers or feasibility as acutely as for example the right to subsistence can be argued to have.

For a state, the duties correlating with the HRFE can be divided into three main types. Firstly, the positive duties to avoid interference with and to protect the arena for the HRFE. This duty includes protecting citizens from harm that may come from expression. This duty will be central in the discussion of Mill’s view on FE in the next chapter. Secondly, the negative duty

[2] Shue's conception of basic rights correlates them with duties. His claim is that these duties are at least threefold (in a simplified version), needing different roles and aspects of duty and organization. If an individual would claim a right in a «state of nature» it would be sufficient to demand the other to refrain from active harm, but not make reference to compliance with a social structure of conditions that might «guarantee» the enjoyment of this right (Shue, 1980, p.38). «Social guarantees against at least the standard threats», that Shue wants duties to secure, require some form of society or government for their realization (Shue, 1980, p.38). And, the demand to the enjoyment of a basic right is normally not secured by the non-interference of individuals alone, but through some social guarantee of protection or securing of the right on a societal or governmental level (Shue, 1980, pp.38-9). Social guarantees might consist of positive actions to guarantee the realization also of negative rights (Shue, 1980, p.39). Then again, social guarantees might also protect against the interference of harmful intereference by third parties. Such protection might protect the autonomy and independence of an individual or group. This protection against likely involves positive measures in the form of legal structures or protecting agencies (Shue, 1980, p.40). On Shue’s account of duties, the positive versus negative rights dichotomy does not hold up in a clear and meaningful way.

of noninterference with the HRFE.[3] The noninterference includes also the Hohfeldian incident of not having a claim against someone practicing their HRFE to not do so. The positive duty of enforcement of noninterference is part of the protection of the arena for the HRFE.

The HRFE creates most clearly duties of the state towards citizens. But the HRFE is a universal right, meaning it is a right pertaining to every person (UDHR, 1948). To ensure universality, citizens' negative duty of noninterference with each other's HRFE seems apparent.

Communication of expression in many cases needs both an author and an audience. It seems that the HRFE provides for the rights of both these parties. If a playwriter writes a play that is banned from theaters, both the author's expression and the audience's access to it are implicated. But for an expression to exist, it doesn't need both an author and an audience. If a government prohibits people to look out of windows out of fear for knowledge of some kind, it is the people with windows whose HRFE is implicated; the view from the window is not an author. It seems that if there is an author who gets censured, this always implicates the HRFE (in the sense of a right to communication) of both parties, regardless of whether the potential audience is aware of the loss of access to communication or not.

Alexander points out that the HRFE "is best thought of as belonging to the audience", but as a right to noninterference only, not a claim right to be provided with communication of some sort (2003, footnote 1, p.41). Also, it concerns communication that would commonly be accessible. Natural interference with communication is not interference the HRFE protects against. So, if a novel gets published in Japanese only, this is not a breach of the HRFE for people who cannot read Japanese.[4]

1.2 The legal content and scope of the HRFE

Article 19 of the Universal Declaration of Human Rights [UDHR] (1948) proclaims that:

> Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers. (UDHR, 1948).

[3] Philip Pettit claims that freedom as noninterference is not enough. Freedom as nondomination is real freedom (Pettit, 2011). But, according to the Hohfeldian analysis and also Mill's account of the HRFE, noninterference is sufficient to ensure a liberty.

[4] Duties to accommodate accessibility do exist correlating with the rights of persons with disabilities. Article 21 of the convention on the rights of persons with disabilities recognizes duties on the part of the state to enable access to the HRFE in a manner that is equal to access enjoyed by persons without disabilities. Thus, the law does not consider disability a natural interference with human rights (Convention on the Rights of Persons with Disabilities, 2007).

The HRFE is thus a Hohfeldian claim right. Every person must be able to enjoy, with the claim to noninterference, the liberty of holding opinions and express them if they wish to. This includes the right to a free access to others' expressions and the free exchange of ideas and knowledge. Also, the HRFE is inclusive of all media (i.e. all communication methods, devices or services) and not subject to (state) frontiers. All forms of communication and language are included in the HRFE.

The HRFE has been used and adapted in other human rights instruments and covenants. Looking at these provides some insight into interpretations of its scope and function. Of central concern, according to legal sources like the Strasbourg Human Rights Court, is the crucial role FE plays in a democratic society. A democracy requires the plurality of voices and opinions present in a society to be allowed their expression. As the Strasbourg Court claimed («Handyside v United Kingdom» App. 5493/72, 1979-80), it is important that the HRFE not only includes content like majority opinions, inoffensive expressions, and ideas experienced as uplifting or trivial, but also and importantly those «that offend, shock, or disturb the State or any sector of the population» (Rainey, Wicks & Ovey, 2014, p.436).

Recognizing the power of expressed ideas to influence, inform, mislead or harm individuals raises questions regarding the scope of this liberty. Our paradoxical intuitions about FE include both feelings of indignation when this right gets violated in any way that seems blatantly «wrong» to us, as well as an instinctive urge to censor expressions we experience as being too extreme and counter to our values or moral intuitions (Alexander, 2003, p.39). With a diversity of opinions and intuitions about what should be in- or excluded from the HRFE comes the question of defining and justifying the HRFE's scope.

The noninterference clause in the HRFE seems to contradict the possibility of an absolute right to FE. A right to noninterference correlates with (at least) someone's duty to refrain from interfering. And this duty may constitute a limitation on the duty-bearer's HRFE. But this problem is not specific for the HRFE. All rights are limited in this way. Rights need an account of how they relate to each other, of how to solve problems of prioritization or the weighing of rights against each other. This is an important topic, but not essential for the discussion of FE on social media and therefore beyond the scope of this thesis.

The HRFE does not proclaim restrictions on the right's scope, but implementation of the HRFE in other conventions and national legal systems establishes regulation of what content

can be legitimately expressed, depending on context.[5] The European Convention on Human Rights refers expressly to FE's correlated «duties and responsibilities» and legal restrictions:

> 1. Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers. This article shall not prevent States from requiring the licensing of broadcasting, television or cinema enterprises.
> 2. The exercise of these freedoms, since it carries with it duties and responsibilities, may be subject to such formalities, conditions, restrictions or penalties as are prescribed by law and are necessary in a democratic society, in the interest of national security, territorial integrity or public safety, for the prevention of disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others, for preventing the disclosure of information received in confidence, or for maintaining the authority and impartiality of the judiciary (ECHR, 1950).

Content of case-law from the European Court of Human Rights confirms that FE intended in Article 10(1) includes expression through any medium and in any form, such as printed media, filmed media, information online, and spoken statements on any medium (Rainey, Wicks & Ovey, 2014, pp. 435-6). Article 10(2) aims to establish ground pillars for the scope and legal restriction of the right to FE. It refers to duties correlating with the liberties. Also, it states the core interests a democratic society has in legally restricting the HRFE in service of its necessity for democratic legitimacy. So, both the scope of the right and its correlating duties and restrictions are defined in relation to the democratic structure of a sovereign nation.

The content or substance of what a person wants to express may be expressed without it being a claim right. But, having the HRFE adds the essential dynamic of being able to claim its enjoyment from the correlated duty-bearer. Pettit points out that a benign dictator might be committed to noninterference with expressing views and thoughts (Pettit, 2011). A benign dictator may even provide extra support for these expressions to be effective, by providing a platform and audience and freedom to follow up. But this would be provided the dictator allows this. There would not be a secured right to the FE to do so.

[5] In Norway, "Grunnlova" (Norwegian for "the constitution") proclaims that freedom of expression does not allow for expressions that express threats, slander, invasion of privacy, continuing harassment, or extreme pornographic, discriminating, or hatefull expressions (Kierulf, Gisle & Elden, 2018).

The HRFE plays an essential part in the practice of other human rights as claimable rights. Without the freedom to express one's opinions, and importantly one's claims and liberties, other claim rights and liberties miss an essential component of their communicability and claimability. Therefore, the HRFE can be considered as "a touchstone of all the other freedoms to which the United Nations is consecrated" (Hannikainen and Myntti, 1993, p.276). Therefore, the General Assembly implies from its very start in 1946 that the HRFE is basic to other rights and liberties declared in the UDHR.

Summary

The HRFE commits us to, firstly, universality, as it says, "Everyone has the right to freedom of opinion and expression" (UDHR, 1948).

Secondly, to freedom of opinion, freedom of expression, and the free flow of information and ideas, as it says, "this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas" (UDHR, 1948).

Finally, that the human right exists concerning all of media and any geographical location, as it says, "through any media and regardless of frontiers" (UDHR, 1948).

In the next chapter I will provide the normative philosophical account of FE that will enable a discussion of the HRFE's relationship with social media. The function of FE that I discuss is its role as a protector of liberal democracy.

2 The HRFE as a protection of liberal democracy

Freedom of expression as a human right applies universally, which means that it is not contingent upon the form of government of a society. Therefore, it is not necessarily the case that HRFE needs for a society to be democratic. FE is still a human right and of great significance if a society is governed by a tyrant. But, nevertheless, a powerful argument for the HRFE can be given by pointing to its relationship with democracy. Throughout this thesis I argue from the view that the HRFE is essential for protecting democracy.

In this chapter, I first provide an account of liberal democracy. I follow David Miller's account of liberal democracy because it is a general, descriptive, and uncontroversial explanation. It will provide some basic understanding useful for the discussion that follows.

Then, I give the traditional liberal defense of FE based on John Stuart Mill's account. Important concerns are how FE relates to and protects democracy, and what constitute legitimate grounds for restricting FE.

According to Mill, the only legitimate reason for interfering with FE is to protect from harm (1989, p.13). This is the harm principle. Ronald Dworkin and Jeremy Waldron are two contemporary philosophers who represent two different views on the role of the harm principle in liberal democracy. These three authors all work within the liberal tradition, but they have different conceptions of harm in liberal democracy, and therefore of the scope of FE.

I end the chapter with five conditions that constitute the traditional account of FE, for it to be protected as a human right, insofar as it is fundamental for keeping a liberal democracy. Based on what Mill, Dworkin, and Waldron have in common in their views, all three think of society as "offline" society and this is reflected in the five conditions.

2.1 A basic conception of liberal democracy

Democracy is most roughly thought of as "government by the people" (Miller, 2003, p.48). This can be a direct democracy in which the government literally is the people. But, in our age, representative democracy is most common. In a representative democracy, political authority belongs to a body of delegates elected through a majority vote. The role of the citizens is legitimizing (and in a way delegating) political authority through electing who may represent them (Miller, 2003, p.38). Of course, a citizen may herself choose to be electable for governing functions.

A liberal democracy as a political system determines the role, limit, and legitimacy of political authority. According to Miller, the democratic process of legitimizing political

authority rests on two basic premises (2003, p.38). Firstly, that persons are naturally equal, meaning that any inequality (i.e. hierarchy of power or authority) must be accounted for and legitimized. Secondly, the view that the good of the citizens is best entrusted to the citizens themselves. This premise is realized through making the people the ultimate source of political authority. Thus, anyone who gets delegated a political power is "accountable to the people as a whole" (Miller, 2003, p.38).

Most liberal democracies reduce the political power and role citizens have in governing to three specific actions. Firstly, the right to vote at elections. Secondly, the right to vote in a referendum. (Referenda can often make or influence a major, possibly constitutional, decision by majority vote). Thirdly, the possibility of lobbying issues of specific interest to persons or groups of persons (Miller, 2003, p.40). Of course, in addition, a citizen's possibility of running for office is also part of their political power.

One obvious weakness of democracy is the imbalance in influence between majority and minority groups in society. Lobbying activities give minority groups the possibility, through the quality of specific action, to somewhat make up for their lack of quantity in relation to majority opinion. Another way in which minority groups and individuals can affect majority opinions is through voicing their opinions and openness to dialogue. Dialogue needs for all parties to be honest and open to counterviews, to be listening to others, and hopefully for the majority to account for the minority voice in some way. Importantly, respect and equality afforded through dialogue is a way for members of a current majority group to acknowledge the possibility of being in the minority position at some point in time (Miller, 2003, p.52-3).

2.2 The traditional liberal defense of freedom of expression

John Stuart Mill's argument for FE is still (despite being from 1859) the most widely used philosophical theory on FE. And this is the defense of FE I will be using in this thesis.

Mill's account of FE builds on his normative view on what it means for an individual to live an ethical life in society in relation to a government that guarantees civil liberty and protection from harm. FE presents a citizen with both the right and the duty to participate in dialogue and contribute to a diversity of views in the democratic fabric of society.

Mill's idea of a liberal society is an extension of his utilitarian but also perfectionist ethics and is richer and more demanding than what one commonly thinks of as a liberal society. His

emphasis on self-improvement, honesty, and perfectionist ethics comes from his belief that a liberal democracy works best with citizens who want to realize their potential in this manner.

But Mill's account of FE has two levels. The perfectionist view is the highest level of human potentiality in liberal democracy, in the form of a process of self-enlightenment. The second level is the minimum account of FE in liberal democracy, namely FE as part of civil liberty with only the basic condition of no harm to others. On both accounts, FE is meant as a protection of liberal democracy.

In this section I first present Mill's view on utilitarianism and perfectionism. Then I present Mill's liberal defense of FE. Five central conditions for FE in Mill's account are that: expression is not harmful; implicitly, that one person counts as one voice; the openness to dialogue of honest speakers; real and morally accountable people; and liberal democracy being a closed, physical society. These five conditions allow for and require the greatest possible scope of freedom of expression.

The only legitimate interference by the government with FE as a civil liberty is to protect others from harm. I finish the account of the traditional liberal defense of FE with a discussion on the harm principle. Because Mill left the principle so vague, there exist very different interpretations of its scope.

Ronald Dworkin and Jeremy Waldron represent two influential contemporary voices in the discussion between libertarian and egalitarian concerns about FE, specifically regarding the question of what constitutes harmful expression, and whether we should restrict the FE to protect against this.[6] I mention them to show how "harm" gets interpreted differently within the liberal tradition.

Mill's view on utilitarianism and perfectionism

Mill's account of FE provides a normative foundation for the conditions of FE as a protector of liberal democracy. The context for *On liberty* is Mill's utilitarian view, i.e. the aim of achieving the greatest possible happiness for the maximum number of persons. Mill considers utility the fundamental principle of an ethics in support of human beings as progressive and truth-seeking beings (Mill, 1989, p.14).[7]

Mill's views on utilitarianism and liberalism are grounded in his perfectionist ethics (Donatelli, 2006, p.163) (Brink, 2018). To Mill, perfectionism is an expression of an inner

[6] Although Dworkin would not have described himself as a libertarian, his views are closer on this point to libertarianism.
[7] Liberal democracy does not need utilitarianism. It works just as well with deontological - or virtue ethics.

transformation of the self. Perfectionism is a process in which a person notices a lack in their present inner state, like for example a belief that is lacking (renewed) experience. Ideally, there follows a self-directed process towards a greater understanding and realization of perfection.

To comply with an outer expectation instead of one's individual impulse would deny the perfectionist process (Mill, 1989, p.67) (Donatelli, 2006, p.162). For Mill, the dignity of human beings lies in intellectual development, i.e. in the questioning and intellectual capacity that comes from following conscience and reason (1989, pp. 35-6). Its opposite lies in the "mental despotism" of oppressive common believes that may not be questioned (Mill, 1989, p. 36). Self-examination and practical deliberation are therefore two central capacities serving the individual and thereby society's progressive understanding of truth (Mill, 1989, pp.66-7) (Brink, 2018).

Philosopher Martha Nussbaum argues that Mill's perfectionist ethics strike an enriching balance between the utilitarian account and Aristotle's idea of happiness and the good life (Nussbaum, 2004, p.62). Benthiam hedonistic utilitarianism unwillingly includes the possibility of ethical perfection in the form of an evil doer who gets pleasure from doing harm (Nussbaum, 2004, p.63). Aristotle equals the good life with happiness, rather than pleasure, and thereby escapes the problem of pleasure from malicious intent. Aristotelian happiness comes from acting on the excellence one has realized as a human being and does not necessarily include pleasure at all (Nussbaum, 2004, pp.64-5). Mill strikes a balance between Bentham's utilitarianism and Aristotle's happiness from human excellence. While Mill recognizes the utilitarian view on pleasure and pain as defining ingredients to the good life, he adds the Aristotelian necessity of human dignity to be realized in the striving for ethical excellence (Nussbaum, 2004, p.66).

Perfectionist ethics also drive Mill's views on liberal democracy. Liberal democratic society is envisioned as "a society […] in which people driven by a constant urge to find and realize themselves would wish to live" (Donatelli, 2006, p.163). Utilitarianism places determining value on an individual's pleasure, but perfectionism recognizes that desire can be manipulated and falsified by external expectations or social norms that reflect the tyranny of the majority (Donatelli, 2006, p.163). Thus, Mill's accounts of utilitarianism and liberalism are deepened by a perfectionist view on how liberalism needs to protect the individual's striving for the good life against the excessive influence of government or society's majority opinion (Donatelli, 2006 p.163).

To further understand how Mill sees the role of FE in liberal democracy, it is necessary to understand his argument for a progressive understanding of truth.

FE in service of society's progressive understanding of truth

Rather than striving for the realization of a fixed idea of truth and what "the good" consists of, Mill believes that human dignity lies in the collective pursuit of a dynamic and progressive understanding of truth. To this end, society should protect the liberty of opinion and expression (Mill, 1989, pp.15-6 and 20-1). FE, then, has an instrumental value for collective truth seeking.

Dialogue and discussion are necessary aspects of a well-functioning Millian FE. Because, even if we believe an opinion we hold to be completely true (which people have a natural tendency to do), we should respect the notion that "if it is not fully, frequently, and fearlessly discussed, it will be held as a dead dogma, not a living truth" (Mill, 1989, p.37).

Our understanding of a truth is dependent on how we weigh arguments and objections and on what grounds (Mill, 1989, p.38). The opposition to our current view should be sought out in its most convincing form, preferably coming from a person who completely believes this counterview (Mill, 1989, p.38). Discussion of a view facilitates an understanding of the reasons one has for holding the view and turns it into a real force in one's mind (Mill, 1989, p.41 and 43). The alternative, neglect of questioning our believes, may lead to holding them as views we say we hold, while our thoughts, feelings, and actions reveal that we actually live by something quite different. This alternative, Mill believes, is the state of the majority of people (1989, p.42).

Following Mill's reasoning, then, it seems that minorities who express their views may be in the best position to succeed in perfectionist ethics. Because, with the human tendency to be convinced of the truth of our opinions, and in addition, most people not actively seeking out opposition to views, this means that the majority of people with majority views do not gain or retain the level of powerful impression a belief should have on one's mind and imagination in the way Mill believes is necessary. Minorities are more likely to meet counterviews regularly, invited or not, and may thus more often partake in the kind of dialogue Mill's ethics prescribe.

Mill claims the same about the aliveness and strength of new teachings still conquering a place in society. Their argumentation gets challenged frequently and fiercely by the majority view, serving the power of the idea on the intellect and actions of the dedicated individual.

This is the true power of dialogue that Mill believes FE should facilitate. It is not just the right to express oneself without interference, but also a duty (both of the individual and the government) to seek truth and to keep our current opinions alive and open for debate, in service of a progressive understanding of truth and the actualization of our human dignity as intellectual beings. Mill's ethics thus serve to realize our human potential as individuals and as a society. Freedom of opinion and expression provide us with the necessary means.

Mill does not believe that truth is more powerful than falsehood. Truth needs to be free from the persecution of those who oppose it for it to find its rightful place in society (Mill, 1989, pp. 30-1). Therefore, truth needs FE to protect the necessary contention of prevailing views. Legal systems should not penalize the expression of any opinion, since this creates and reinforces the "social stigma" of a certain belief or opinion (Mill, 1989, p.31 and 34).

Threats to liberty, according to Mill, can come both from the citizenry or from the government. Citizens can pose a threat in the form of the tyranny of the majority; and government can rule with its own form of tyrannical power (Mill, 1989, pp.7-9, 19). FE in Mill's conception is a safeguard against the tendency of both majority groups and governmental forces to tyrannize others by forcing their opinion on them. FE therefore protects both democratic equality among citizens and the validity of democratic relations between government and citizens. I will explain how Mill perceives these two threats of tyrannical power.

The tyranny of the majority consists of an almost inescapable social control. According to Mill, this form of tyranny restricts people in subtle and far reaching ways by permeating public discourse and social life with society's dominant morality (1989, pp.8-9). The individual needs some form of protection against being undermined and restricted in this way. In a democracy, the tyranny of the majority is the mishandling of power that Mill is most worried about.

The tyrannical power of government lies, according to Mill, in when it exercises its power from an assumption of infallibility (1989, p.22). Since Mill sees truth as a dynamic, progressive phenomenon, it thrives on a diversity of opinions. A diversity of opinions seeking new understandings and perspectives requires FE. To disrupt this process through censorship or baseless claims of the falsehood of specific opinions, is an illegitimate assumption of infallibility (Mill, 1989, pp.20-1):

> To call any proposition certain, while there is any one who would deny its certainty if permitted, but who is not permitted, is to assume that we ourselves, and those who agree with us, are the judges of certainty, and judges without hearing the other side (Mill, 1989, p.25).

The problem with an assumption of infallibility is thus not the fact that a person, a majority, or the government has an opinion on what is true or right, but that one claims the power to decide upon the matter for others (Mill, 1989, p.26). This would censor certain opinions or

aspects of truth and thus limit the scope of FE. But, Mill claims, no person can ever be legitimately prevented from expressing their opinion (1989, p.20).

> […] the peculiar evil of silencing the expression of an opinion is, that it is robbing the human race; posterity as well as the existing generation; those who dissent from the opinion, still more than those who hold it. If the opinion is right, they are deprived of the opportunity of exchanging error for truth: if wrong, they lose, what is almost as great a benefit, the clearer perception and livelier impression of truth, produced by its collision with error (Mill, 1989, p.20).

Mill's argument from truth then, is that every opinion contributes to society's progressive understanding of truth over time.[8]

From this discussion it can be concluded that an important condition for FE is no interference with expression, provided the expression is not harmful. This is the first of five conditions of the traditional account of FE. The other side of the coin of harmless expression, is the legitimate interference with expression to protect from harm. I now turn to the question of when interference with FE is legitimate.

Legitimate interference with expression to protect from harm

According to Mill, FE can legitimately be restricted if the content and/or the circumstances of the expression cause harm to others. Mill's view is referred to as the harm principle, and is the claim "[t]hat the only purpose for which power can be rightfully exercised over any member of a civilized community, against his will, is to prevent harm to others" (Mill, 1989, p.13).

Addressing FE directly, Mill says that interference with expression is legitimate if the situation causes "their expression a positive instigation to some mischievous act. […] The liberty of the individual must be thus far limited; he must not make himself a nuisance to other people" (Mill, 1989, p.56). Mill gives the following example to illustrate his point:

> An opinion that corn-dealers are starvers of the poor, or that private property is robbery, ought to be unmolested when simply circulated through the press, but may justly incur

[8] I don't know what Mill grounds his belief in a progressive understanding of truth on. Mill seems to think it (close to) impossible to know whether we are right or wrong. His argument for enlivening our opinions on truth instead of letting them be empty dogma's does not account for these opinions' truth. Mill's view on a progressive understanding of truth poses epistemological dilemma's that I will leave aside. So I assume, with Mill, that progressing towards truth is possible.

> punishment when delivered orally to an excited mob assembled before the house of a corn-dealer, or when handed about among the same mob in the form of a placard. Acts, of whatever kind, which, without justifiable cause, do harm to others, may be, and in the more important cases absolutely require to be, controlled by the unfavourable sentiments, and, when needful, by the active interference of mankind (Mill, 1989, p.56).

We may therefore interfere with freedom of speech if it causes some harm to others or if it leads others to commit harmful acts. When the harm in an expression comes from the circumstances it is expressed in, interference is only indirectly a restriction of the HRFE, i.e. not a restriction on its content. This explains why Mill thinks that in the corn dealer case, it is acceptable to print the same content in a newspaper. The intention is not to stop the expression, but to prevent harm. But, the published opinion in a newspaper may effectively be a similar instigation, just in a less directly traceable way. It is hard to know where exactly to draw the line on restriction.

Political theorist Johann Go addresses several "contemporary challenges" to Mill's account of FE (Go, 2018, p.3). Go argues for a reevaluation of FE considering these challenges but is committed to Mill's account as being the most convincing account of FE (Go, 2018, p.5). I agree with Go that Mill's account of FE faces new forms of challenges, but I think that Mill's account of the principles supporting FE is applicable to these challenges and remains valid.

According to Go, Mill's conception of harm should be understood within Mill's socio-historical context (Go, 2018, p.7). Mill's argument for FE via print, but for a restriction of the same content expressed near its object, such as in the case of the corn dealer, illustrates that Mill did not foresee the effect of media in contemporary society.

In addition, Go points out that Mill excluded psychological and emotional harm from his conception of harm:

> [H]is reasoning for this is primarily empirical and based on the resultant effects. If hate speech has effects beyond mere emotional harm such as causing the person to commit suicide, or if the hate speech has the effect of inciting violence, it may then fall under the purview of the Harm Principle. Mill also refers to harm in numerous other forms throughout On Liberty, including the violation of interests, the infringement of rights, and the instigation of "mischievous acts" (Go, 2018, p.8).

So, for Mill's harm principle to account for a conception of harm that would be acceptable to contemporary society, we must leave behind Mill's outdated views on psychological and emotional harm. Go argues that "as our empirical knowledge about harm improves, this affects the scope of the state's ability to regulate expression under the Harm Principle" (2018, p.9). As research on psychological processes and long-term consequences of experiences advances, a contemporary conception of harm may include things (like bullying) that Mill typified as mere "psychological offence" of insufficient consequence to count as harm (Go, 2018, pp.9-10).

So, though the harm principle provides Mill's account of FE with a clear principle, there is a challenge to its applicability. One needs a clear conception of what constitutes harm. In part, this also means there needs to be agreement on where to draw the line on an expression being a traceable or immediate cause for harm.

Ronald Dworkin and Jeremy Waldron each give a different conception of harm that one can have as a liberal. Consequently, each suggests a different scope of FE and gives an argument for what legitimizes this scope. I will start with a discussion of Dworkin's account and end the section with Waldron.

Dworkin: principle over utility

Ronald Dworkin (1931-2013) was a professor of law and philosophy who wrote extensively on rights and legal philosophy, including influential titles like "Taking Rights Seriously" (1977). Dworkin's argument for FE is that a society cannot legitimately expect or force any citizen to respect laws or majority decisions when they have been prohibited from partaking in society's democratic and moral fabric leading up to it (Dworkin, 2009, p.viii). This can be called his "legitimacy argument for FE".

Like Mill, Dworkin sees FE as a necessary condition for the insurance of a government's equal obligation towards each citizen, the best political organization of which is a liberal democracy (Dworkin, 2006, p.132). Unlike Mill, Dworkin believes that society's conception of FE should be justified by it as a principle, not by its utility (Dworkin, 2006, pp.130-1).

Dworkin claims that Mill bases his argument for FE on the beneficial consequences it provides (2006, p.130). But, according to Dworkin, the justification for liberty and thus also for the HRFE must come from principle (2006, p.131). FE as a principle is grounded on human dignity. Human dignity, according to Dworkin, is a condition which needs active recognition by the government:

> We can find [the HRFE's basic principle] in a condition of human dignity: it is illegitimate for governments to impose a collective or official decision on dissenting individuals, using the coercive powers of the state, unless that decision has been taken in a manner that respects each individual's status as a free and equal member of the community (Dworkin, 2009, p. vii).

It is human dignity that grounds both an individual's claim to be a free and equal member of society, as well as the duty of the state to treat each citizen accordingly. According to Dworkin, dignity is unconditional, based on a normative idea. This seems equivalent with the UDHR's claim that human dignity is inherent in every human being (UDHR, 1948). Both agree on dignity being a status that each human being has that grounds certain human rights and its correlated duties.

Dworkin agrees that a society should protect members of minorities that are vulnerable to harmful consequences of hate and prejudice. But, according to Dworkin, such protection should not happen on the level of FE. Rather than limiting the scope of FE to protect from harmful speech, a government should implement laws that achieve minority protection without reducing FE (Dworkin, 2006, p.132). Dworkin warns against limiting FE as it would fragment a principle that he claims needs to be indivisible for it to maintain its power (2009, p.ix).[9]

The democratic fabric of society needs to guarantee the legitimacy of its outcomes. This democratic fabric is constituted of several factors. Every competent person of age has the right to a political vote. Also, each person has FE. This implies the possibility of expressing one's ideas, not necessarily to influence others but at the very least as an equal participant in society's moral and political processes (Dworkin, 2009, p. vii). Any democratic majority decision has legitimacy only if in the process towards this decision no one has been excluded from the possibility «to raise a voice in protest or argument or objection before the decision is taken» (Dworkin, 2009, p.vii). In this manner, democratic processes optimize the premises for being representative of the entire electorate, confirming each person's dignity by their equal right to participation. Not restricting democratic participation before landing on a majority decision, is, according to Dworkin, how FE legitimizes liberal democracy.

[9] "It is tempting, […] to think that even if some liberty of speech must be counted a universal right, this right cannot be absolute; that those whose opinions are too threatening or base or contrary to the moral or religious consensus have forfeited any right to the concern on which the right rests. But such a reservation would destroy the principle: it would leave room only for the pointless grant of protection for ideas or tastes or prejudices that those in power approve, or in any case do not fear. We might have the power to silence those we despise, but it would be at the cost of political legitimacy, which is more important than they are. Any such reservation would also be dangerous. Principle is indivisible, and we try to divide it at our peril" (Dworkin, 2009, p.ix).

Dworkin's view is supported by Mill stressing the necessity of contestation for assumptions of truth. As Mill says,

> Complete liberty of contradicting and disproving our opinion, is the very condition which justifies us in assuming its truth for purposes of action; and on no other terms can a being with human faculties have any rational assurance of being right (Mill, 1989, p.23).

We need to be contested and challenged in our opinions for us to be as sure of their correctness as is humanly possible. For Mill, FE as the freedom to contest and be contested in all opinions and beliefs, serves the individual and governmental duty to attain the greatest degree of honesty and truth one is capable of (Mill, 1989, p.22).

Dworkin might be handing Mill a solution for the challenge of providing a useable conception of harm. Dworkin's post-election government has the democratically legitimized authority to regulate society to protect (minority) citizens from harm. All the while, the full capacity of FE to democratically represent each citizen is secured as an inviolable principle. On Dworkin's account, the harm principle does not diminish the scope of FE.

But Dworkin's FE as a protection of the legitimacy of democratic outcomes, can be argued to have problems exactly in relation to its effect on democratic processes.

Philosopher Eric R. Boot, as an element of his duty-based approach to ethics and human rights, points out difficulties with "limitless freedom" (Boot, 2017, pp.147-150). To provide context for his argument, Boot refers to Benjamin Constant's distinction between the conception of freedom in ancient political philosophy versus what can be called the modern understanding originating with Hobbes. According to the ancient understanding of freedom, citizens had the freedom (the "positive liberty") to *partake* in the state's direct democracy.[10] This positive liberty gave citizens a "social power" in relation to the community and the state (Boot, 2017, p.147).

[10] In ancient Athens, direct democracy consisted of the practice of political speech in front of the assembly. Therefore, the political structure was practically synonymous with its core values of isegoria and parrhesia. Isegoria and parrhesia are two different free speech principles. Isegoria is "the equal opportunity of […] citizens to speak in the principal political institution of the democracy, the Assembly" (Werhan, 2008, p.300). Any citizen could speak and be heard on any of the matters that the assembly decided on. Parrhesia is the practice for citizens "to speak openly and frankly once they had the floor" (Werhan, 2008, p.300). Parrhesia was central for a real and "authentic public debate among citizens who honestly and forthrightly spoke their minds" (Werhan, 2008, p.317).

In contrast, Hobbes' definition of liberty was "no stop, in doing what [one] has the will, desire, or inclination to [do]", which included protection against civic duties (Hobbes, 1996, p. 149). According to Boot, this modern conception of freedom is a negative liberty, meaning one's freedom is protected against interference and the possibility of duties restricting it. This protection is guaranteed through rights such as the HRFE (Boot, 2017, p.147).

Dworkin's view on individual rights as a concept is that they should be inviolable. When an individual has a right "then it is wrong for government to deny it to him even though it would be in the general interest to do so" (Dworkin, 1997, p. 269) (Dworkin referenced in Boot, 2017, p.147). Thus, Boot points out, on Dworkins' account individual rights trump the common good. This view is taking liberal democracy to mean that the sphere of individual liberty takes priority over public considerations and in principle has no correlating civic duties (Boot, 2017, p.148). It is this one-sided and limitless conception of freedom, in which it is the sole focus, trumping all other concerns, viewing duties to be infringements on the right to freedom, that Boot argues has problematic consequences.

When rights are limitless and trump all other concerns, citizenship is reduced to a conception of "what the community must do for us", which according to Boot leads to a "passive and detached understanding of citizenship" (Boot, 2017, p. 148). Passivity comes from viewing citizenship as essentially consisting of being a rights-holder. Detachment occurs because of viewing oneself in our citizenship as essentially unrelated to the community (Boot, 2017, p. 148). Boot argues that we should have a conception of civic duties *in addition to* a conception of human rights (2017, p. 148). This means that there should be some public- or civic concerns that do affect the scope of human rights like the HRFE.

Boot claims that active civic participation needs encouragement or enforcement as a duty. Mill would agree that civic participation in the form of dialogue and honesty are moral duties enhancing the potentiality of FE. Also, that FE's utility means that its scope should be restricted in line with the harm principle.

But I think Boot's objection against Dworkin does not do Dworkin justice. Dworkin's FE cannot be said to trump the common good as an individual limitless right, if we consider that the correlating civic duty lies in accepting and abiding by the democratic majority outcome consequent to the HRFE. Civic duty, on Dworkin's account lies in ceasing the opportunity to influence society democratically by using one's FE and then accepting the outcome as legitimate and representative of the common good. The utility of legitimate restriction of freedom comes after the democratic outcome.

Dworkin and Mill don't share Boot's solution for concerns of social cohesion in a liberal democracy. But should the harm principle take into consideration egalitarian concerns of civil liberty? This consideration is addressed in Jeremy Waldron's view on freedom of expression, which I discuss in the next section.

Waldron: protecting people from hate speech

Professor of law and philosophy Jeremy Waldron (1953) argues for the regulation of FE to protect human dignity in the form of equal social status of citizens in a liberal democracy (2012, pp.4-5). Waldron focuses especially on the protection from hate speech of those who are a member of minority groups that historically have had a vulnerable position in society (Waldron, 2012, p.5).

Like Mill and Dworkin, Waldron argues for the centrality of FE in liberal democracy. But Waldron argues for the inclusion of hate speech in the conception of harm that society should protect against. He prioritizes social equality over individual liberty in determining the scope of FE.

Waldron's concern is that hate speech damages the inclusiveness of a community and the confirmed equality of its members. Inclusiveness and equality are common goods that Waldron thinks a liberal democracy should endorse and protect (2012, p.4-5).

Central to Waldron's argument is his view that hate speech undermines and attacks human dignity. Dignity lies in the security of a «basic social standing» as «proper objects of society's protection and concern» (Waldron, 2012, p.5). So, dignity is an actively supported property of an integrated and inclusive society, rather than an inherent property of human beings independent of recognition or the society they are part of. Following Waldron's argument, dignity is something that can be gained or lost, which makes it contingent on social recognition. Therefore, human dignity is vulnerable to prejudice, hate, and exclusion, and needful of protection against hate speech.

Acknowledging the harm that can come from speech, Waldron argues for restricting FE when it comes to hate speech. Waldron's argument focuses specifically on hate speech as «part of the permanent visible fabric of society» (2012, p.3). Visible expressions of hate such as posters, graffiti and other visual material, visible and present in a community, affect the community through conveying a twofold message. Firstly, it sends the message of denigration, unwelcome, and a threat of unsafety to a specifically targeted group within the community (Waldron, 2012, p.2). Secondly, it aims to inform others with similar hateful views that they

are not alone, that they belong to a larger group of likeminded people. This gives these peers power and confidence in their opinion (Waldron, 2012, p.2-3).

Waldron's point of status as a social vulnerability seems legitimate. The question is what kind and how much hate speech has the power to accomplish this. Who is justified in determining this, and how does Waldron weigh these concerns against the individual liberty a liberal democracy is meant to protect?

Discussions concerning what legitimizes hate speech regulation have revealed what Brink calls "libertarian and egalitarian strands within the liberal tradition" (Brink, 2001, p.119). Waldron is exemplary of the egalitarian stand on liberalism, whereas Dworkin can be said to represent the libertarian stand.

The egalitarian line of thinking stresses the divisiveness and harm hate speech causes and supports restriction of FE on this basis (Brink, 2001, p.119). The liberty of individuals and members of minority groups is reduced when the basic dignity and equality of personhood is challenged by discriminatory and hateful speech. Therefore, the egalitarian line of thought argues that, "[t]hough one might well be reluctant to restrict speech, it might seem that the correct response to hate speech, as with other forms of discrimination, is regulation" (Brink, 2001, p.119).

Libertarian reasoning does not deny the ugliness of hate speech and its effects on people and society, but it believes restricting FE is harming our core liberties, which is at least as harmful (Brink, 2001, p.119). More than (almost) anything, freedom needs to be protected, and the cost of doing so is noninterference with expression, also with the expression of hateful and repulsive views. Instead of restricting speech, the libertarian solution for hate speech is using FE to counter the hateful views. The libertarian view is that egalitarian concerns should be addressed in this manner, through more expression rather than restriction (Brink, 2001, p.119).

Waldron might reply to this objection from the libertarian stance that using FE to further one's views does not address the damage hate speech does in the process of countering it. Thus, the libertarian and egalitarian stance disagree on whether it is harm to our core liberties or harm to equality of personhood that needs to be prioritized when considering the legitimacy of restrictions on FE.

According to Mill's view, minorities need to be ensured the equal access to civil liberty, an essential civil liberty being FE, and protection from the tyranny of the majority. But individuals need not be shielded into equality. Such shielding is, on a Millian account, more likely a slippery slope into the threat of a government's assumption of infallibility since it necessarily builds on assumptions on an idea of truth and its enforceability. Also, the wider the conception of harm,

the greater the restrictions on civil liberty to protect against it. Waldron sees protection from harm as the greater necessity for the securing of a contingent human dignity, where Mill and Dworkin prioritize civil liberty to secure respect for inherent human dignity.

With this I finish the discussion on Mill's harm principle and continue with a discussion of the other conditions implicit in Mill's account of FE.

The second condition for FE on Mill's traditional account of FE is, one person, one voice. Implicit in Mill's account of FE is the assumption that every individual in liberal democracy represents one voice. FE needs for every voice to have equal power. The FE of each voice is further facilitated by the willingness for dialogue between persons with contrary views.

The third condition of FE is the openness to dialogue of honest speakers, and a consequent diversity of opinions. Mill argues that the progressive understanding of truth is served by diversity, and regardless of whether an opinion is true or false (Mill, 1989, p.57). If an opinion is true, those who are in error have a chance of changing their view. If an opinion is false, those who have the truer opinion will gain from the experience of the contrast with the false opinion. And in most cases, opinions are only partly true (Mill, 1989, p. 20 and 57).

Speakers need to be honest and, according to Mill, both the individual and government share a duty of honesty. According to Mill, "[i]t is the duty of governments, and of individuals, to form the truest opinions they can; to form them carefully, […]" (1989, p.22). This feeds into Mill's expectation that "[m]en, and governments, must act to the best of their ability. There is no such thing as absolute certainty, but there is assurance sufficient for the purposes of human life" (Mill, 1989, p.22). So, the liberty of FE serves or contains a duty to be honest and act on our knowledge to the best of our ability.

Honest dialogue gives the best chance of maintaining a degree of truth in public discourse and opinion when we realize that most opinions are not true or false, but are a share of both (Mill, 1989, p.47). Even contrary opinions are, according to Mill, both likely representative of some truth, though in different parts (1989, p.47). In short, all opinion has a valuable role to play in society's collective and progressive search for truth. Censorship is a limitation or halt to this process that cannot be justified (Mill, 1989, p.23). Thus, the widest possible scope of FE (restricted only to prevent others from harm) is needed for a conception of truth to the best of our current collective ability.

The fourth condition for FE is that speakers are real and morally accountable people. According to Mill, accountability is essential for FE. Ten, discussing how Mill related his views to democracy, writes on Mill: «it was precisely [the ballot's] secrecy to which [Mill] objected.

If voters had to account publicly for their vote, they would act more responsibly" (Ten, 1998, p.374). According to Mill, then, secrecy or anonymity erodes the individual tendency to act responsibly. This argument can be extended to FE. Being identifiable when one expresses an opinion will create a sense of responsibility for the expression and in addition ensures the capacity for the community to hold the speaker accountable.

As a fifth condition, the Millian account implies a closed, physical society. FE is one of the rights people have as members of society. The liberalism Mill discusses is the civil liberty of citizens in relation to their government. It gives an account of the purpose and scope of an authority's legitimate power in regulating civil liberty (Mill, 1989, p.5). It does not account for external influences or global virtual relations. Thus, the traditional account of FE has as an implicit premise a conception of society as a closed, physical nationally defined entity.

Closed, physical society and virtual neighborhoods

The fifth condition of the traditional account of FE is, as mentioned, a closed, physical society. Waldron's account illustrates this condition, since, though it does mention the internet occasionally, it fails to address how essential differences between real physical neighborhoods and online viral reality affect FE.

Waldron argues that debates on restrictions on hate speech should focus on expressions that are part of our lasting and visible environment, rather than on the comparatively short-lived spoken word (Waldron, 2012, pp.37-8). Waldron's focus on the visible environment makes expressions into something physical, a society's physical environment and reality.

> [T]he sort of attacks on vulnerable minorities that elicit attempts to regulate and suppress "hate speech" include attacks that are printed, published, pasted up, or posted on the Internet – expressions that become a permanent or semipermanent part of the visible environment in which our lives, and the lives of members of vulnerable minorities, have to be lived. No doubt a speech can resonate long after the spoken word has died away […]. But to my mind, it is the enduring presence of the published word or the posted image that is particularly worrying in this connection (Waldron, 2012, pp.37-8).

In arguing for prioritization of hate speech restrictions on printed or otherwise visible expressions of "hate speech", Waldron treats as essentially equivalent an image posted on a neighborhood wall and an image posted on Facebook. The examples he mentions have in common that he considers them a (semi)enduring part of the visible environment in which we

live. Pointing out that it is the environment we all live in, he implies the inescapability of meeting these expressions, likely again and again. Waldron's believes hate speech restrictions should "stop [visible expressions of hate] from becoming part of the landscape, part of the evident stock of people's ideas circulating in a society and looming over the environment in which people live their lives" (Waldron, 2012, pp.148-9). Though he is less clear on whether freedom of the internet should undergo the same restrictions as other published expressions, he is unifying in his conclusion that hateful messages should not be allowed as part of the environment in which we live. I think Waldron means that "the environment in which people live their lives" includes both our real and virtual neighborhoods. I think that, for his argument concerning hate speech, Waldron is not looking for anything more specific than this. The internet then, according to Waldron, is just one of several and essentially equivalent modes of (semi)permanent visible expression that are part of "the marketplace of ideas" (Waldron, 2012, p.176). He does not discriminate between the difference in access to information in one's physical society or an online virtual "neighborhood".

Concerning expression online, Waldron's argument could specify between the treatment of expressions and the potential for effect they have depending on where the expression is published. Waldron is not making a clear separate case for restrictions of expression online based on hate speech but based on his argument he should.

The dynamic between what traditionally constitute majority- and minority groups is different online when it comes to identity building and effect. Waldron's argument for the protection against hate speech is the negative effect it has on the reinforcement of prejudice. Such negative reinforcement happens online in the same manner, but with greater reach and speed. A real-life example of this is the French lesbian girl who received up to two hundred hate messages per minute after having reacted with hate speech to hate speech directed at her from a Muslim person regarding her sexual orientation (Sage, 2020). It is difficult to compare this case with Waldron's example of meeting hate speech on a wall on a stroll through one's neighborhood. But on the other hand, no minority is a minority online. FB's community building capacity facilitates large communities around features of identity (rather than geography). These examples are all examples concerning quantity and quality of impact and context. Although Waldron's argument stands intact when it comes to the kind of effect hate speech has according to him, it is also clear that the potential online expressions have are different from offline expressions.

Does Mill place the bar too high?

A possible objection against Mill's conditions for FE is that his views on human beings and society are too idealistic. His normative account places the bar high when it comes to the moral duty of honesty and the expectation of progress it places on both individual and government. But Mill provides two qualifications of his account that show that, though he is idealistic about the human perfectionist potential, he tries to account for lower and realistic expectations.

One qualification concerns the openness to dialogue of honest speakers. He acknowledges that society's understanding of truth may not be so progressive at all. Since humanity tends to change outdated partially true majority opinions for other (more contemporary) partially true opinions, it does not really progress towards truth but changes opinions in relation to the age's morality (Mill, 1989, p.47). To Mill it is important that dialogue with honest interlocutors has the greatest likelihood of ensuring that when society moves away from one partially true view, it at least moves to another partially true view and not to a false view. Though, on a Millian view, even a majority experience with complete falsehood most likely would enable a development towards some truth if society maintains the FE to make this possible. In this manner, Mill's conditions for FE seek to maintain the dynamics that tend towards progress and truth.

The other qualification Mill provides is that the perfectionist individual likely always be a rarity. "In sober truth, […] the general tendency of things throughout the world is to render mediocrity the ascendant power among mankind" (Mill, 1989, p.66). According to Mill, when a democracy has no influence of "a more highly gifted or instructed one or few", it is the naturally mediocre majority that will lead to a mediocre democracy (1989, p.66). Progress can only come from the individual and it is this potential in each human being that civil liberty must protect and facilitate. Mill acknowledges that most people will not realize this potential. The gift of the average person to society is his ability to recognize and be inspired by a "highly gifted or instructed" individual (Mill, l989, p.66).

Thus, it seems an unfair misrepresentation of Mill to see him as proclaiming the view that liberal democracy is a collective of honest and ethically perfectionist individuals that contribute to a steadfast collective progression towards self-realized and actualized truth. Rather, Mill's view seems to be that human dignity consists of the individual potential for self-realized progress. Society and government do not naturally encourage this individual potential, but rather tend to facilitate a mass mediocracy. Despite this, for society to protect, allow for, and be served by the human realization of individual potential for truth and excellence, liberal democracy is the best form of governance. To ensure liberal democracy, we need FE.

2.3 Five conditions for freedom of expression

A liberal democracy is the political form Mill's liberalism envisions, with special caution towards the threats of abuse of power. Such power abuse can, according to Mill, take the form of a moral tyranny of society's majority, or tyrannical powers active in government expressing itself in the form of censorship and/or assumptions of infallibility. Truth, and the freedom of expression necessary to sustain and obtain it, are thus expressive of and protective of liberal democracy.

The traditional account of FE, for it to be protected as a human right, insofar as it is fundamental for keeping a liberal democracy, has these conditions:

1. The expression is not harmful. Harmless expression cannot legitimately be moderated or tailored. FE on a traditional Millian account requires the greatest possible scope, only legitimately interfered with to protect from harm, where what is harmful is decided through democratic deliberation. This is the basic condition for FE and determines the scope of FE as a claim right.
2. One person, one voice. In a liberal democracy, each individual is a physical person who counts for one voice.
3. The openness to dialogue of honest speakers. The interlocutors intend to be honest and intend to engage in dialogue. This creates the diversity of opinions that Mill thinks necessary for realizing a human being's potential and for society's progressive understanding of truth.
4. Real and morally accountable people. Interlocutors are real and identifiable people. This ensures people being morally and legally accountable.
5. A closed, physical society. The context that is implied in accounts of FE and its legitimate restriction from Mill, Dworkin, and to a large degree Waldron, all imply society being a closed and physical society.

Mill's understandings of society and liberal democracy and the related form of FE do, however, imply that the FE's conditions seem stronger than what may be obvious for contemporary society. Firstly, for Mill, speakers are *real* speakers. A voice belongs to a person and is not anonymous. Secondly, Mill calls upon the honesty of a speaker. Though his argument does include reference to the value of false opinions and the likely mediocracy of the majority, it is rooted in the premise of a speaker's moral duty to strive for truth. Thirdly, the HRFE as Mill understands it does not just allow for, but requires dialogue, the willingness to hear the other side. Lack of engagement with or openness for others' views or opinions makes for an

impotent democracy. Plus, Mill's analysis is of a closed society. But contemporary society has external influences in our political debates and international relations, and from social media.

The question is whether the worldviews inherent in the philosophical views I discussed have the same validity in relation to online reality. Mill could not have foreseen the communication made possible by social media. Surely, Mill did not conceive of "disembodied" comments in real-time from effectively anonymous persons who could be anywhere in the world. These kinds of expressions were not part of his conception of society and its progressive search for truth through the free exchange of expressions and ideas. But that Mill did not foresee current forms of online communication, does not necessarily mean that his view on freedom of expression does not account for this. Whether it can or not will be an important question when discussing FB as a case in the next chapter.

Even Waldron, whose book is fairly recent, builds his argument on an extremely physical and geographically located conception of society. And does Dworkin's inviolable FE account for the dreadful extremes brought forth by online platforms that factually do allow for complete freedom of expressions in a way offline liberal democracy doesn't? 8chan, being one of the few platforms that host non-moderated, unrestricted expression, has turned into a platform that promotes harm in its ugliest extremes (Roose, 2019). It is hard to see how unrestricted expression that facilitates child pornography networks or terrorist support networks through online communication can be an argument for the protection or legitimacy of democracy through FE.

So, if it is the case that Mill, Waldron, and Dworkin argue from the premises of an offline form of society and governance, do these sufficiently resemble premises that apply to the realities online? Are their arguments valid for our current free speech situation? Comparing a traditional conception of FE with the online society social media represent might give us new insight into how to approach old questions in new relevant ways.

In the next chapter, I present how FB works and whether expression on FB fulfills the conditions of the traditional account of FE. If FB does, then FB does not violate or challenge our conception of the HRFE. If FB does not, then the follow-up question is what this conclusion means for the HRFE in the age of social media.

3 How Facebook makes the traditional account of FE run into trouble

In this chapter I describe aspects of Facebook that correlate with aspects of the traditional account of FE. In the previous chapter I discussed how the traditional account of FE can be summarized in five conditions for it to be protected as a human right. In this chapter, I will explain in five sections in what way FB correlates with each of these five conditions. I end each section with a short discussion. I aim to show that there are important discrepancies between the traditional account of FE and expression on social media like FB.

In this thesis, when I refer to "Facebook" I mean the social media platform by this name. I write "Facebook Inc." when I refer to the business that owns several social media platforms (a.o. Facebook, Instagram, WhatsApp). But mostly I will focus on the structure and content moderation specific for the social media platform called Facebook.

Mark Zuckerberg, founder and CEO at FB, and others at FB have said and written a lot about FE, the principle of "voice" and FB's decisions on content moderation. Though they likely do not intend their words to be philosophical principles or arguments, I will nevertheless distill some principles and arguments implicit and explicit in their claims.

I first present some information on FB's business statistics before starting the discussion of aspects of FB that concern the traditional account of FE.

FB has, according to statistics provided by them, as of June 2019, 2.41 billion users that are active on FB at least once a month (Facebook, 2019a). At that same time, they have about 39.650 employees worldwide, and an average amount of 2.1 billion users that use one of their social media platforms (Facebook, Instagram, WhatsApp, Messenger) *on a daily basis*. For comparison, the most populated country in the world, China, had in 2019 a population of 1.38 billion. So, the population of the largest country in the world equals only roughly 65% of Facebook Inc.'s daily users.

These statistics give an indication of the enormous portion of online expressions that FB hosts. FB Inc. moderates expressions of 27% of the world's population every day.[11]

[11] 7.7 billion people in 2019.

3.1 The expression is not harmful

Freedom of expression concerns harmless expression. Mill claims that there is no legitimate ground for interfering with FE unless to prevent harm to another person. This is the basic condition for FE and determines the scope of FE as a claim right.

In this section I illustrate how FB challenges the traditional FE's conceptions of no interference with harmless expression, and also the traditional restriction of expression in line with the harm principle based on democratically justified conceptions of harm.

First, I present FB's principle of "voice" as FB's equivalent of FE. FB aims at maximizing "voice", meaning providing as many people as possible with access to online expression. But FB's business model is based on modifying and interfering with user expressions to collect user data, which in turn creates the company's revenue. The interference with expressions on FB is executed by algorithms and results in individually tailored voices online. Because of this, FB users have no symmetrical access to content on FB in the manner a person has access to information in physical society. Furthermore, user expressions and access to its intended audience are distorted in a way a FB user has no control over.

I discuss Mill's harm principle, which on the traditional account endorses the widest possible scope of FE, only legitimately restricted to protect others from harm. A Millian conception of harm that the harm principle should protect against is, real, direct, and (most likely) individual harm.

FB self-regulates harmful content based on its own conception of harm. A core problem with FB's self-regulation is that, while it should be the democratic community that decides what counts as harm, FB determines their own conceptions and policies.

Furthermore, a challenge specific for FB is how to handle the harm they make possible. Since the platform does not edit expressions before they are published, harm is always only discovered after publication, and the platform spreads content fast with great reach.

FB's format and business structure also facilitate a more abstract nonphysical kind of harm, one that violates liberal democratic procedure such as elections. The individually targeted advertising that constitutes FB's revenue, has been used to target FB users on a large scale with distorted information on political candidates, manipulating voters with proven effectiveness.

FB's "voice" principle

Zuckerberg identifies FB's core principle as "voice" (Zuckerberg, 2019a and 2019b). "Voice" means the opportunity for individuals to communicate or express themselves online.

Zuckerberg has a conviction "that giving everyone a voice empowers the powerless and pushes society to be better over time" (2019b). He illustrates this claim of empowerment and progress by sharing his experience of how the early version of FB transformed campus life of the university he was attending. Zuckerberg believes that the students' voices on FB resulted in that the students

> organized more social events, started more businesses, and even challenged some established ways of doing things on campus. It taught me that while the world's attention focuses on major events and institutions, the bigger story is that most progress in our lives comes from regular people having more of a voice (Zuckerberg, 2019b).

Based on his interpretation of FB's impact on campus, Zuckerberg correlates voice with social cohesion, progress, and democracy, as progressive processes at a grassroots level. He suggests that because of a platform like FB, people will express and engage themselves more in other aspects of life and that the shared platform increases inclusion and social cohesion. From this, Zuckerberg distills two slogans that communicate FB's values. FB intents to "give people *voice"* and create inclusion by "bringing people together" (Zuckerberg, 2019b).

In the introduction to FB's Oversight Board Charter, the importance of voice is explained in a way that can give the impression of alignment with the HRFE: "Freedom of expression is a fundamental human right. Facebook seeks to give people a voice so we can connect, share ideas and experiences, and understand each other", it says (Facebook, 2019c, p.2). It seems that the intention behind the principle of voice here is to acknowledge the HRFE as being of fundamental value. But as I will argue and illustrate in this chapter, despite this impression, FB's Oversight Board Charter does not implement the HRFE, nor do any of FB's other policies.

Still, FB's argument for giving people a voice is an argument for free expression as being essential to society:

> [G]iving everyone a voice is a positive force in the world, increasing the diversity of ideas shared in public discourse. Whether it's a peaceful protest in the streets, an op-ed in a newspaper or a post on social media, free expression is key to a thriving society. So, barring other factors […] we lean toward free expression. It's core to both who we are and why we exist (Allan, 2018).

FB argues that FE is essential to diversity in public discourse, which in turn is a necessity for a "thriving society". FB does not define what is meant by "society" (society could be meant related or unrelated to nations, it could be referring to the community of FB users worldwide, either in reference to or independent from their geographic location), nor what constitutes a thriving society (or a not-thriving one). But what is clear is that "giving everyone a voice" is seen as a good.

I now discuss how FB interferes with user content through the application of an algorithm.

FB's algorithm, creating an individually tailored user experience

FB's public profile is that of a social media platform providing connection as a service to their users, free of charge. FB facilitates connection with for example other users, networks built on common interests, and provides businesses with a customer base. This is reflected in their public slogan of "bringing people closer together".

FB's business model is based on the monetization of data collection, or "surveillance-based advertising" (Cyphers, 2019). FB creates individual user profiles from the personal data it collects. Data collection makes it possible for clients to buy advertisement on FB that targets individuals of their audience (specific to their commercial or political campaign) with great accuracy. It is the high specificity of individualized targeting of FB users that makes it attractive to invest in FB as a commercial platform (Frontline, 2018) (Mosseri, 2016). In this manner, it is the collected user data that create FB's revenue. Personal data has surpassed oil in being the most valuable asset on the market (The Economist, 2017) (Amer & Noujaim, 2019).

In order to optimize and sustain economic growth, FB Inc. is focused on attaining a progressive quantity and quality of data from their users. To achieve this, user engagement is stimulated through the application of algorithms. FB's algorithms manipulate a user's FB experience based on the data already collected, and encourage further user engagement (Mosseri, 2016). The science behind this process comes from research data on the addictive properties of behavior (Amer & Noujaim, 2019).

FB regularly changes the algorithm that determines a user's news feed (Mosseri, 2018) (Zuckerberg, 2018). The implementation of a new algorithm reprioritizes the visibility and ranking of posts on a user's news feed. FB individually tailors each user's experience to maximize the matching of content with data on their preferences (Mosseri, 2016). In effect FB regulates the reach and accessibility of user expressions on its platform and determines what online content a user is presented with.

FB compiles different categories of interest per user.[12] FB does not give insight into how exactly these categories are established. Probably they are based on the data it is known they collect, namely, what posts or groups a user "likes" on FB, what content a user interacts with on FB, what third-party websites and apps a user visits and uses outside of FB, and the physical location of a user's electronic devices (Cyphers, 2019).[13]

Political campaigning is an example of individually targeted advertising that FB facilitates on its platform. As Britanny Kaiser, Cambridge Analytica's former employee who worked for Trump's presidential election campaign in 2016 explains, FB gives the best engagement rate on P.R. investments, therefore FB gets the biggest share of the advertising budget in campaigns (Amer & Noujaim, 2019). Cambridge Analytica strategically focused their resources on a group of FB users they called "the persuadables". The persuadables were voters whose mind they thought they could change. In Trump's presidential election campaign, the persuadables were users with specific personality traits who lived in the "swing-states". Kaiser explains that persuadables are bombarded with psychologically manipulating info until they are likely to see the world the way the campaign wants them to, assuming they then act accordingly. In this case, the persuadables were manipulated into voting for Trump (Amer & Noujaim, 2019). Individually targeted content thus facilitates psychological manipulation with the purpose of changing a person's behavior. The techniques used are developed by the military for psychological warfare (Amer & Noujaim, 2019).

Having a voice on FB thus creates the company's revenue. Besides the interference from tailoring each FB user's voice, FB's algorithms hinder the symmetrical access to information.

Symmetrical access to information

The HRFE includes the right to access to information. In a physical society such as implied by the traditional account of FE, access to information is symmetrical among individuals. This means that when Joan reads a copy of the Washington Post, in Washington D.C. and John reads his copy of the Washington Post in Berlin, both will have access to the same content. This symmetrical access does not exist on FB. Joan's news feed will be moderated into an optimal representation of her established preferences, and John's FB newsfeed will be individually tailored according to his. Neither of them has access to the real unmodified content on FB.

[12] Much of a user's data is visible in a user's personal "Ad Preferences page" on FB.
[13] When someone "likes" a post, to the public this is a sign of connection, approval, or some other form of communication. To FB this same "like" is a piece of personal data (Frontline, 2018). FB collects data from online behavior of their users, both on and off their platform (Frontline, 2018).

Moderating online content can be argued to be a violation of the HRFE as the right to access to information, since interference with expressions on the traditional perspective only is warranted by the harm principle.

User control over content

At the user level, connections on FB are established by taking positive actions. FB users can send another FB user a friend request that needs to be accepted by the other part to establish a connection. One-side-initiated connections are the possibility of "like"-ing and following sites from businesses or organizations, joining open groups (versus both-parties-dependent connection of FB-groups with membership approval by a moderator), or active browsing of other users' news feeds that give universal access (i.e. also to those who are not FB-"friends").

The control a user has over barring certain content from their news feed lies in the possibility of "(un)following", "blocking", and "hiding" content (Mosseri, 2016). Of course, friends can be "unfriended" and liked pages can be "unliked", removing these connections from a user's network. These actions are done without consent from or informing the other party. All personalized controls are used by FB as data on how to further predict, rank, and tailor content (Mosseri, 2016).

These actions constitute a user's conscious control over its connections and preferences on what people and topics are represented on the user's newsfeed. But the range of these choices and the content that appears is subject to FB's algorithm and policies, over which a user has no control and of which very limited knowledge. A user cannot choose to have access to the complete unregulated online content on FB.

FB's algorithm is in effect a continual content moderating loop outside of a user's control and awareness. It is artificial intelligence [AI] that works from a research-based assumption about what user expectations are, i.e. what it is that keeps a user sufficiently satisfied to return to and engage with content. The engagement with content provides FB with new individualized data that gets fed into the algorithm. The algorithm then further ranks and manipulates available news feed data to increase personalization of the news feed. Interaction with this content then further determines individualized manipulation of content, etcetera. The algorithm thus effectively takes each individual FB user further and further away from a direct meeting with FB content as it is. The likelihood of meeting exceptions to preferences are reduced over time and the natural diversity of expressions on FB is inaccessible to the individual.

FB's surveillance based advertising and concurrent algorithm are problematic for the HRFE because they interfere with what content reaches a user and a user's control over who they reach

with their expression online. Furthermore, each user gets presented a distorted representation of actual online expression. FB is a holographic universe in which each user is presented with only their individually tailored image of the world. Despite FB's efforts to give users the feeling of being in control, in practice this control only gives access to possibilities of network building. It does not ensure access to online expressions as they really exist.

FB's handling of harmful content

According to Zuckerberg, FB's aim is to discover and handle harmful content as soon as possible, mainly using AI, and preferably before anyone reports it (2019b).[14] What Zuckerberg means by handling of this content depends on the category and the user who posted it.

Based on FB's policy on the spread of harmful information, the main FB-page of a well-known former footballer was removed for repeated violations (Quinn, 2020). The user spread harmful conspiracy theories, sometimes linked with anti-Semitic hate speech. But while the removal of the page was a clear gesture enforcing FB's policy statements, the problem was not solved completely. Videos spread by the user still appeared elsewhere and had over a 30 million views; and besides the deleted page with some 770,000 followers, another page with more than 68,000 followers was still up (Quinn, 2020).

False content, as seen in the Covid19 related examples, is only removed if it is likely to cause physical harm to persons. In 2018, this policy applied only to content from "countries where there is ongoing conflict", the plan was for it to "later be rolled out globally" (Facebook, referenced in Kozlowska, 2018). The Covid19 pandemic, then, either introduced the policy globally or is treated as a conflict.

False content causing physical harm may include, then, the instigation of physical harm that also Mill refers to in the case of the corn dealer. On FB, likely also physical harm consequent to false medical advice. Other less harmful content gets flagged or marked by FB to warn users of potential triggers or offense.

In its Community Standards, FB recognizes that expression and communication online have a different capacity and scope from (public) discourse offline. It says, "[o]ur commitment to expression is paramount, but we recognize the internet creates new and increased opportunities for abuse" (Facebook, 2019b). An example of the speed and reach of harmful content and threats on social media is the case of a 16-year-old French girl who stated on Instagram that

[14] FB has developed a system to mark what it deems harmful content, consisting of about twenty categories (Zuckerberg, 2019b). Some of these categories can be inferred from FB's Community Standards page. FB's definition of harmful content includes nudity and some forms of hate speech.

Islam is a religion full of hate. The girl received death-threats, also, information about where she went to school was posted online (BBC, 2020a). As a result, she left Instagram, left school and stayed at a hide-out address until threats against her were investigated and contained. She claims that at one point she received 200 hateful responses per minute (Sage, 2020). Human nature has not changed due to social media, but the capacity for harm may have increased because of it.

In addition to increased reach, online technology creates new forms for the expression of content. Before social media platforms, sharing a picture would mean physically showing or mailing a picture to another person. Online, images are shared almost instantly, with a person at almost any geographical location. FB's technology for live streaming can, in a positive sense give people the opportunity to participate in important events (Zuckerberg, 2019b). But the same technology can facilitate participation in or being victim to harm. Examples of the live streaming of harmful events include the mass shooting at a mosque in Christchurch, New Zealand, and self-harm or suicide attempts (Kelly, 2019) (Begley, 2017).[15]

To diminish the consequences of the live-streaming of harmful content, FB employs AI-systems that aim to detect harmful content as soon as possible (Zuckerberg, 2019b). But the nature of live-streaming makes it impossible to prevent the streaming of harmful content altogether.

The problem of direct publication affects all FB's handling of harmful content. FB's current format of direct self-publishing by users means the handling of harmful content will always happen after the fact. And with the potential reach of online content, this means the content will likely have reached some audience before it is moderated or removed.

FB's self-regulation of harmful content

FB, as a self-regulating company, determines its own policies on harmful content. These policies build on FB's self-defined conception of harm. Harm regulations get intertwined with FB conditions for what counts as true or false content. FB users are limited in their online expression and access to other's online expressions to content which FB defines as *sufficiently truthful in the right way*, and its correlative *not harmful in the wrong way*. Only "harmful" content gets removed, but many gradations of harm or falsehood are moderated, tagged, or

[15] After a wave of criticism in 2017 of FB's permissive policies on the live streaming of self-harm, Zuckerberg now says FB is especially concerned with the danger of self-harm among young people (2019b).

semi-censored. Thus, content that is not recognized by FB as being true and harmless is distorted in its online presentation by FB as a content moderator.

FB self-determined policies on content moderation seem to reflect arbitrariness, rather than stable principle. Misinformation for example, on many occasions gets limited exposure on the platform. As Zuckerberg points out, there are problems connected with regulating and defining a category like misinformation (2019b). Hoaxes are the clearest example of misinformation he gives, and these can be tagged to make them recognizable for AI. But satire and exaggeration are problematic to recognize for AI systems. And media expressions that represent a certain limited angle on a factual occurrence pose another problem (Zuckerberg, 2019b).[16]

In addition, FB has exceptions to its policy on misinformation. Political advertising and newsworthy content are not fact-checked:

> [w]e don't do this to help politicians, but because we think people should be able to see for themselves what politicians are saying. And if content is newsworthy, we also won't take it down even if it would otherwise conflict with many of our standards (Zuckerberg, 2019b).

So, the category of misinformation that is restricted seems to contain: hoaxes, possibly some degrees of satire and false exaggerations, some media expressions. Exceptions are (i.e. content which does not get fact checked or moderated) political advertising and newsworthy content. But these again have exceptions: "Of course there are exceptions, and even for politicians we don't allow content that incites violence or risks imminent harm - and of course we don't allow voter suppression. Voting is voice" (Zuckerberg, 2019b).

In an attempt to reduce the examples to premises guiding FB's misinformation policy, we might say that FB aims to regulate any content that is not factually true, taking into consideration conversational modes that naturally contain degrees of falsehood (i.e. satire, exaggeration). Exceptions are expressions FB deems important for public discourse and likely its revenue, meaning political advertising and newsworthy content.[17] The whole segment of

[16] Zuckerberg says, "Take misinformation. No one tells us they want to see misinformation. That's why we work with independent fact checkers to stop hoaxes that are going viral from spreading. But misinformation is a pretty broad category. A lot of people like satire, which isn't necessarily true. A lot of people talk about their experiences through stories that may be exaggerated or have inaccuracies, but speak to a deeper truth in their lived experience. We need to be careful about restricting that. Even when there is a common set of facts, different media outlets tell very different stories emphasizing different angles. There's a lot of nuance here. And while I worry about an erosion of truth, I don't think most people want to live in a world where you can only post things that tech companies judge to be 100% true" (Zuckerberg, 2019b).

[17] Addressing objections, Zuckerberg says: "I know many people disagree, but, in general, I don't think it's right for a private company to censor politicians or the news in a democracy. And we're not an outlier here. The other major internet platforms and the vast majority of media also run these same ads" (Zuckerberg, 2019b).

misinformation seems subject to the overarching restriction of "harmful" content (incitement to violence, risk of imminent harm, nudity). And then Zuckerberg adds a consideration related to the protection of FB's core principle of "voice".

Even if there are clear guiding principles to FB's content moderation, they are not apparent from how they communicate their decisions and policies.

Discussion

The traditional account of FE is committed to no interference with all harmless expression. FB's business model based on the trade on FB users' data results in interference with the online voice of each FB user.

FB offers its users an account on its platform free of charge. Most people understand that online services that seem to be free of charge are not actually free services but monetize their services in some other way. My guess is that most people think they pay for online services by having adds pop up at inconvenient moments. Ads can be ignored or clicked away, so they are a small price to pay. But on social media the user does not pay by exposure to ads, but with the valuable asset of personal data.

Aware consumers know that they pay for services by giving up control over their personal data. What users cannot know with complete certainty, however, is when and how data is harvested, who has access to it and to what use it is put (Amer & Noujaim, 2019). What is known is that data harvesting for FB users exceeds what a user publishes on the platform. It includes private messages sent on FB's private messenger service and tracks online movements also when a user is logged off from the FB platform.

The harvesting of personal data by social media is not a main topic for the question my thesis addresses. But the consequences of the ways in which personal data are harvested by social media companies strongly affect a user's FE through the related content moderation processes that distort each individual news feed.

Zuckerberg resists using traditional views on expression to evaluate online expression. He considers online speech a form of expression in its own right (Zuckerberg, 2019b). He claims that the internet is too different from past communication platforms to legitimize using historical precedent for reference.

Zuckerberg believes that what makes the internet to be distinctly new is the quantity of voices online (2019b). Nearly half the world population uses social media. One might argue that (almost) the entire world population has always had a voice in some way or another. But

having an online voice means one has access and reach in a way that did not exist before the internet. According to Zuckerberg, this large online population empowers people and stimulates the FB values of social cohesion and progress (2019b). But the empowerment Zuckerberg mentions does not include an individual's power over one's own expression in the way that the HRFE is meant to secure.

FB's algorithmic tailoring of online voices creates three forms of interference with harmless expression. Firstly, users have no control over the distortion of the information one publishes or has access to on FB's platform. A user cannot choose to "turn content moderation off" or moderate it to include a more realistic diversity of expressions in one's news feed. Nor does a user have access to information about how the content on their newsfeed is manipulated.

Secondly, the confirmation of an individual's tastes and opinions creates "echo chambers", i.e. the preferences a user puts out online are what comes back, giving an artificially created experience of constant confirmation of one's views. Thus, algorithmic moderation interferes with a user's access to a natural diversity of voices and opinions.

Thirdly, FB users lack symmetrical access to information and expressions. Since FB individually tailors every newsfeed, there is no realistic representation of FB content present anywhere on the platform.

But one could argue that there is no interference with an individual's FE from FB since users know that their user experience is individually tailored. If a person dislikes this, they can leave FB. They can join other social media or access the information through another medium. It would only be interference if they had no other means of accessing information.

I think this objection fails to recognize FB Inc.'s prominence in contemporary society. Choosing not to take part in any of FB Inc.'s self-regulated online services, means not having an account on FB, Messenger, WhatsApp, or Instagram. Effectively, this leaves a user with the choice between Twitter or a Chinese driven social media platform like TikTok. Additionally, given how many persons do use the platform, the impact of social media like FB on liberal democracies should still be addressed.

Besides the algorithmic tailoring of user content, FB's handling of harmful content is an important aspect of how the company fails the traditional account of FE. FB's practice as a platform for expression when it comes to the prevention and restriction of harmful content faces two objections when held up against the Millian account.

The first objection concerns FB's capacity for being an adequate duty-bearer. When it concerns self-harm, AI systems detect a risk and contact the person with information about

relevant first responders. An obvious criticism of FB's system is the use of machines rather than people to reach out to a person in distress (NRK, 2020). This concern may seem practical and relatively detail oriented, but it is illustrative of a more structural problem. The problem FB faces is whether they can realize the duty they have acquired by facilitating a seemingly limitless capacity for world-wide real-time expression.

If we grant that FB is not accountable for the content that is expressed on its platform, what is it FB is accountable for? FB is a tech company that created the online platform for its users' instant expression. FB is also the actor that manages the platform and makes a profit from doing that. As the facilitator and manager of the expressions on its platform, FB seems logically accountable for the facilitation of its potential for harm and has a (moral) duty to manage this as part of its general management. But does FB have the capacity to realize this duty?

FB tries to regulate harm mostly in two ways. It uses AI and user alerts to detect harmful content as soon as possible, and it uses AI and user alerts to detect fake accounts as a source of harmful content as soon as possible. Both methods need to deal with staggering amounts of alerts and are, because of FB's format, always at least one step behind online reality. An argument in support of FB is that the company is trying to realize the duty. But this leads us to the second concern.

The second objection against FB's way of dealing with harm concerns FB's self-regulated and non-transparent way of conceptualizing harm. FB should acknowledge and practice democratically defined conceptions of harm and legitimate regulation of harmful expression. This could take the form of national conceptions of the international agreement on the HRFE. But instead, the company abides by its own ongoing determination of what counts as harm without providing users or outsiders with clear information about their current conception and the related practices. FB does not relate in a clear and accountable way with their users and the world outside their company concerning how they conceptualize and regulate harm.Though FB acknowledges a role as a duty-bearer regarding content management, they avoid relations of accountability that a duty correlates with.

In self-regulating harm, FB seems to get stuck in attempts to sort out phenomena on its platform, rather than defining core principles to base its content moderation policies on. Focusing on handling phenomena gives FB policies the impression of arbitrariness.

FB's principle of "voice" seems to function as FB's equivalent of FE. But FB's "voice" has no defined role or identity that grounds its content moderation. When considering all that FB says about its principle of voice, it is not clear what this principle's goal or direction is. FB

gives no cohesive account of online expressions, and consequently lacks an account of online expression in relation to online and offline society.

FB seems not to have a clear conception of the principles it claims to have, nor are principles Zuckerberg deems “preferable” necessarily in any way implemented in FB’s policies or business structure. For example, Zuckerberg mentions regularly that “[a]s a principle, in a democracy, I believe people should decide what is credible, not tech companies” (Zuckerberg, 2019b). But this “principle” is not identifiable in FB’s practice since users cannot decide on content moderation or censorship.

The fact that FB defines and acts on its self-regulated incongruent conception of (truth and) harm, plus the fact that there do not exist any avenues of action for users to affect this conception or decide on FB’s policies, are problems when holding FB against the standard set by the harm principle as part of the traditional account of FE.

In addition, there seem to be different kinds of harm online. Harm on FB can be traced back to two factors: FB users and FB’s business structure. FB users cause harm by publishing harmful content. Such harm can come in the form of for example malicious misinformation, often posted through anonymous accounts. Or harmful content can come from identifiable users who are honest about their extremely hateful views. These forms of harmful expression are easiest to align with the traditional account of the harm principle.

By using the corn dealer example, Mill makes clear that an essential aspect of the harm principle as it regards FE is, that it should not actively instigate harming others (Mill, 1989, p.56). Aspects of the harm principle, then, are that individual liberty can legitimately be restricted when harm is done to others, or when there is active instigation of harm to others. A propaganda campaign spread on FB by Myanmar military against the Rohingya minority in their country, resulted in “murders, rapes and the largest forced human migration in recent history” (Mozur, 2018). On social media, malicious misinformation with the intent of manipulating its reader into an action such as participation in the Rohingya massacre, thus are harmful instigations.

The harm caused by FB's business structure is more abstract and of a different kind. One of the downsides of FB's sales of individually targeted advertising, is that it facilitates what can be called "harm done by third parties". This is not the kind of real and direct physical harm that the traditional harm principle refers to, but harm to democratic procedure. With the way social media are organized, some claim that there are currently no valid democratic elections possible (Cadwalladr in Amer & Noujaim, 2019) (House of Commons, 2019).

But is it legitimate to call this phenomenon of undermining democratic procedures "harm"? And how does this phenomenon relate to FE? An example of abuse of FB's sales of individually targeted advertising is when Cambridge Analytica targeted the persuadables with tools of psychological manipulation into voting Brexit in the UK, voting for Trump in the USA, and similarly manipulated many other democratic elections (Amer & Noujaim, 2019).

Is this phenomenon harm? It is not uncommon for companies to psychologically influence or manipulate people into certain behavior. Advertisement is an accepted form of exactly this. But Cambridge Analytica used the psychological manipulative techniques in a manner and on a scale normally used only in military psychological warfare (Amer & Noujaim, 2019). Thus, it can be argued that it is the scale that makes the phenomenon harmful.

But one may object that the scale does not alter the principle. So, if we allow for advertising, we should allow voter manipulation in this manner.[18] Then this online phenomenon is not harming individuals. Furthermore, companies have a claim right to the HRFE just like individuals have (Nelson, 2020). The scale of the expression's targeted audience does not change the HRFE of Cambridge Analytica, nor the HRFE of the campaigns that hired Cambridge Analytica.

The next question, then, is whether harming democratic procedure is a form of harm to liberal democratic society. Only if the harm principle can reasonably include harm to *society*, can the online phenomenon be called harm. This seems hard to argue for based on the Millian harm principle.[19]

But the solution could be found outside FE, in a different perspective on the matter. The Cambridge Analytica case has revealed, after governmental research, the inadequacy of current election law in dealing with the impact of social media on campaigning and the democratic process (House of Commons, 2019). The matter of distorting election outcomes might thus be delegated to election law.

[18] Though Mill points out, there is no FE without the concordant freedom of individual autonomous action to express and ensure diversity of opinion (1989, p.57). Others' customs, norms, or opinions, expressed in a way meant to affect or restrain individual autonomy, according to Mill, both make unattainable individual happiness and "individual and social progress" (1989, p.57). For Mill, human worth is an expression of the exercise of all our human faculties to develop and maintain the intellectual and moral capacity one has (Mill, 1989, p.59). Manipulation of consumers or voters thus violates human worth as autonomous individuality. Thus, liberal democracy is violated in it being expressive of the democratic collective voice of society consisting of autonomous individuals. But this Millian argument is more in line with his perfectionist account than with the more basic account of FE and the harm principle.

[19] To make this a legitimate view, liberal democratic society needs to somehow be an individual actor, it seems. Furthermore, it seems to require an account of what the equivalent of direct physical harm is for society as an entity.

But what role *does* the HRFE play? Under current legal conceptions of the HRFE, FB Inc. is a right-holder with the HRFE (Nelson, 2020). And it is governments who carry the (main) duty of regulating society to ensure the HRFE and protect right-holders from harm, not FB. But, as I will show in this chapter, FB's format and business structure impact individuals and liberal democratic society in far reaching and problematic ways.

To once again go back to the Cambridge Analytica example, even if FB users are not harmed by the massive and targeted influence from third parties FB sells advertisement to, FB does distort the individual online reality of their users to make possible the advertisement. By filtering and tailoring information for the individual FB user, FB distorts reality, and this is against what the principle of FE is for.

3.2 One person, one voice

In the traditional account of FE, one voice counts as one voice. One person has one voice, and this is the shared and equal condition of all persons. Accounts of FE like Mill's build on this as an assumed premise since it reflects physical reality.

Online, a user's voice does not reflect singularity, but its expressions' scale is amplified or drowned out by the content moderating algorithm. On FB, the amplification of a voice is a way to increase user engagement and gain user data. Problematic effects of voice amplification are that it tends to promote polarizing voices online since these provoke the strongest user response. Pluss, that the appeal of polarizing content and the communities that are formed because of this, make users vulnerable to manipulation.

On social media, one person can have several voices by creating several FB accounts. This is not a rare phenomenon, and most often serves one of two purposes. The purpose may be related to personal interests, like identity experimentation or personal safety. The second variation is the creation of a coordinated cluster of fake accounts with the intent of psychological manipulation or harm. Such fake accounts can be created by either people or AI. Both phenomena are matters I discuss in section 3.4, on real and morally accountable people.

Amplification and going viral

Engagement with a post beyond a certain threshold signals to FB's algorithm to give the post a boost. The boost artificially amplifies the visibility and reach of the post across a wider audience. This extends distribution, prolongs user engagement and, in some cases, can make a

post “go viral”. Viral content is an online expression, such as a meme, news article, video, or picture that spreads fast across social media through commenting, sharing, and linking.

Viral content can connect people around a cause, create an online movement with global reach and effects on physical society. An example of this is when the hashtag #blacklivesmatter went viral, which resulted in online communities and also offline protest groups in the USA, Australia, Canada, and the UK.[20]

Since content is impossible to contain once it has started spreading, FB aims to regulate what type of content can go viral: “We especially focus on misinformation that could lead to imminent physical harm, like misleading health advice saying if you’re having a stroke, no need to go to the hospital” (Zuckerberg, 2019b). FB is made aware of harmful content either through AI or users notifying FB of disturbing content. As previously mentioned, AI is fast but cannot correctly interpret all kinds of content. It cannot “read” pictures or recognize satire, for example. Notifications coming from users are likely to come relatively late after the publication of the content, so the post will already have spread online. In addition, the non-AI content evaluation is done by human moderators, meaning the process is time consuming and allows for an even larger online spread of the content. For example, when FB closed down an account for spreading harmful conspiration theories, the videos had spread beyond recall and had over a 30 million views on different online platforms (Quinn, 2020).

But the algorithmic amplification creates other problems that the condition of one person, one voice, is meant to protect against.

Threat of the tyranny of the majority from algorithmic amplification

FB does not follow the one person, one voice condition. This finds expression in the algorithmic amplification and limiting of online voices. Besides the objection that the algorithm distorts a natural presentation of online content, it creates two other problems. Firstly, the tendency of polarization. Secondly, the vulnerability to manipulation that ensues. Together, these online processes can reinforce majority opinion into a Millian tyranny of the majority.

The algorithm typically increases the reach and online presence of posts that represent either a popular majority- (for example, “Barack Obama speech: yes, we can”) or an inciting polarized opinion (for example, “Hillary for prison”). Online expressions that communicate fear or anger provoke the most responses (Amer & Noujaim, 2019). Expressions aimed at polarization use

[20] Black lives Matter protests in Ferguson (USA) lead to, amongst others, a shooting incident during which two policemen were seriously injured (BBC, 2015).

this knowledge to their advantage. Polarizing content that induces strong emotions gets confirmed in a hyper-engagement boost, amplifying the most heated or loved content.

As research journalist Carole Cadwalladr points out, algorithmic exaggeration of polarizing content makes online users not only vulnerable to the polarizing phenomenon itself, but also to manipulation by others (Amer & Noujaim, 2019). Cadwalladr has written extensively on FB, its role in the Brexit election and FB's facilitating role for Cambridge Analytica's voter manipulation (Cadwalladr, 2020). Since fear and anger are the emotions that tend to trigger the most online response they can, in combination with FB's algorithmic amplification, function as mass manipulators (Amer & Noujaim, 2019).

FB users' tendency to create connections around shared views and FB's content moderation create fertile ground for the harmful potential of a large quantity of fake accounts disseminating the same propaganda. Rand Waltzman, leader for DARPA from 2010 to 2015, explains the harm that comes from intentionally harmful information on such a large scale. "Filter bubbles" is a term for when people retreat into online communities in which people share one specific view or interest (Frontline, 2018).[21] The one-sidedness of these online communities makes its users easy targets for customized propaganda disseminated through fake accounts. When this is done on a large enough scale, this distortion of information turns into a weapon. "It's the scale that makes it a weapon" (Waltzman in Frontline, 2018).

An example of how this psychological propensity is used for malicious purposes is the abuse of the "black lives matter" hashtag by foreign actors. When the "black lives matter" memes started going viral online, Russian actors created fake black lives matter memes optimized for emotionally triggering FB users. When a proponent clicked on the meme, she was taken to an event page where she was invited to a real-life protest in the USA that, unbeknownst to the user, in reality was organized by the Russian organization. Moreover, the same Russian actor created adversary groups with the meme "blue lives matter", in support of police people. Both groups were then manipulated into a greater polarization by the same actor, resulting in an effective divide-and-conquer strategy (Cadwalladr in Amer & Noujaim, 2019). So, FB's algorithm encourages confirmation of polarizing views, and in addition increases its capacity for harm, both offline and online.

On a Millian view, the enhancement of popular majority opinion can promote the tyranny of the majority. FB communities tend to be created around a commonality of interest or opinion. The amplification of an inciting polarized opinion may support FB's aim to increase user

[21] DARPA stands for the Defense Advanced Research Projects Agency and is part of the American ministry of defense.

engagement. But with it, FB creates support for trigger-sensitive majority opinion driving people to flock together in a related echo chamber, facilitated by algorithmic preference confirmation and community building based on a shared view. FB increases the tendency of politics and society to create segments of people who perceive their interests as a group to be threatened by other groups with a differing view.

In all of this, the physical reality and diversity naturally accounted for in the condition of one person, one voice, gets distorted or lost out of sight.

Discussion

The biggest concern Mill has about liberal democracy is its potential for a tyranny of the majority permeating the social fabric of society. Popular majority opinions can dominate public discourse and social mores in oppressive ways for those who do not share them. A solidly established and well protected HRFE can protect against this threat by maintaining the rightful possibility for individual- or minority opinion to be expressed freely and heard respectfully, grounded in the physical reality of one person, one voice.

The singularity of a voice is a natural reflection of the equal chance individuals should have in a liberal democracy. The capacity for amplification of voices that tend to divide, rather than invoke a discussion of views among equals, supports dominance rather than democracy.

Posts that create strong responses are those that elicit fear or anger in the reader, polarizing groups of people into adherents and opponents, confirming majority- or minority views in a caricatural manner. This can create an oppressive atmosphere both on- and offline. These same posts, since they are triggering to users, have a real potential for going viral. Algorithmic amplification of provocative voices makes online users vulnerable to polarization and to manipulation by third parties. All this shows that FB's algorithmic amplification threatens equality among users that a nonmoderated expression of each singular voice would naturally support.

3.3 The openness to dialogue of honest speakers

The traditional account of the role of FE in liberal democracy stresses the importance of dialogue. According to Mill, for a person to be willing to engage in dialogue with an honest intention is essential for the function of FE in liberal democracy. And as Dworkin points out, FE is not just meant as a right to expression, but also as a facilitator of dialogue between the

full diversity of opinions present in a society. Therefore, FE requires both for individuals to have access to other voices, and also for each person to have the attitude of listening to and engaging in honest dialogue with these other voices.

The format and content moderation on FB do not contribute to a user attitude of openness to dialogue among honest speakers. FB's content moderation creates an online reality in which users do not naturally read or interact with the expressions of users with opposing views, nor does FB's format facilitate informed discussion or the user attitude of openness to dialogue.

In this section I first discuss how FB's content moderation interferes with user access to a diversity of viewpoints as this is a condition for dialogue among them.

Next in this section, I discuss the fact that Millian dialogue requires honest interlocutors. Honesty is a Millian moral duty and is reflected in Mill's argument that FE serves society's progressive understanding of truth. Mill's perfectionist ethics reinforces this argument with an individual (and governmental) duty to express oneself with the intention of being honest. Mill's argument from truth is nevertheless meant to account for falsities and half-truths as well. The only legitimate interference with FE does not come from dishonesty but from the harm principle.

FB policies do not reflect Mill's perfectionist views on honesty. American online companies are, according to §230 of the Communications Decency Act of 1996 (USA), not liable for the user content published on their platform (Bowers & Zittrain, 2020, p.2). FB is not liable for the platform's user content and, assuming the algorithm is without bias, has no preference for a certain kind of harmless expressions.

Willful dishonesty often has a malicious intent. Before the discussion that closes the section, I address how FB does try to minimize injurious content on its platform and has policies to evaluate and moderate content it deems harmful.

Inclusivity about viewpoints

A diversity of viewpoints is a consequence of the diversity of voices online. FB claims to stand for inclusivity of viewpoints on its platform, meaning it does not willfully exclude or promote certain views.[22] This value means that when FB moderates content, it does not do this to advance a certain view, but solely with the aim of tailoring for individual preference and

[22] Research shows that preferences may be (unwittingly) programmed into FB's algorithm. Epstein and Robertson call it the Search Engine Manipulation Effect (SEME). SEME means that online platforms like FB and Google can influence the political landscape or even elections through bias in their algorithms or search engines (Epstein and Robertson, 2015). The fact that FB uses an algorithm to moderate user content, moderation into which only FB has full insight, makes the SEME a real but difficult to detect possibility.

maximizing user activity. As Mosseri argues, FB does this "not only because we believe it's the right thing but also because it's good for our business. When people see content they are interested in, they are more likely to spend time on News Feed and enjoy their experience" (2016).

Any changes to FB's individualized news feed ranking are made in line with FB news feed values. The values are based on research FB has done on people's expectations about their news feed. In line with these research-based facts about expectations, the prioritization of connections with family and friends is central. Furthermore, news feed aims to satisfy individual preferences in relation to information and entertainment. A user's individually tailored news feed is continuously moderated as part of the process with which FB facilitates surveillance-based-advertising.[23]

The value of inclusivity refers to FB's tolerance for diverse content on their platform overall. It does not mean, based on the information about FB's algorithm, that FB safeguards a diversity of content for each individual user. Since the algorithm ranks content to satisfy established preferences of each individual user, a FB user has the exact opposite experience, namely one of confirmation rather than diversity of views. This confirmation is not limited to the level of interest in certain topics but focuses on specific views on a topic. For example, Facebook does not just note that I am a user interested in the category "politics" but determines my political view. Consequently, I will not receive a diverse or inclusive representation of political voices but only those that target my preferred political view.

According to Mill, real discussion among a diversity of views enlivens individual contact with the arguments and the motivation one has for a specific view. Lack of such dialogical challenge leaves a person more vulnerable to having a superficial understanding of one's views and for manipulation from what seem to be peers.

The Brexit campaign on FB is an example of this vulnerability. In 2019, The Digital, Culture, Media and Sport Committee published its report on "Disinformation and 'fake news'", after investigating the interference with the Brexit elections facilitated by tech companies like FB (House of Commons, 2019) (Amer & Noujaim, 2019). The "Leave EU" organization hired Cambridge Analytica to target those FB users who, based on their psychological profile, would

[23] "We learn from you and adapt over time", it says. And, "[w]e're always working to better understand what is interesting and informative to you personally, so those stories appear higher up in your feed", and, "[w]e work hard to try to understand and predict what posts on Facebook you find entertaining to make sure you don't miss out on those", and "we work hard to understand what type of stories and posts people consider genuine — so we can show more of them in News Feed. And we work to understand what kinds of stories people find misleading, sensational and spammy, to make sure people see those less" (Mosseri, 2016). A news feed is thus constantly updated and fine-tuned to a user's views, preferences and expectations.

be most likely to be persuaded into voting for Brexit. Cambridge Analytica used FB user information and targeted advertising to distribute psychologically manipulative memes to engage a democratically crucial group of people into voting for Brexit (Amer & Noujaim, 2019). Had users on FB been established in a format encouraging challenging and investigative discussion of different views, it is unlikely they would have been as vulnerable to this manipulation.

FB's policies on user engagement and the recent experiences with the Brexit campaign on FB show that FB facilitates one-sidedness and confirmation of views, rather than encouraging intellectually challenging and honest dialogue.

Dishonest and harmful user content on FB

Honesty and harm do not correlate. Dishonesty can be quite harmless and honesty very harmful. But on FB, much of content that is removed because it is considered harmful, is misinformation with malicious intent. There is a continuous influence of intentional falsehoods online in the form of fake news (i.e. news that is misrepresented as being factually accurate), deep fakes (i.e. typically video's in which AI technology has edited a new character into the original video. The aim is to let the video pass as authentic), hoaxes (i.e. false stories or scams, usually meant to victimize someone), malignant misinformation and scams.

An example of this is the online presentation of the Covid19 pandemic, which sparked much speculation about its origin and future consequences. Investigations into purposefully misleading falsehood and online extremism by the Institute for Strategic Dialogue revealed a spike of Covid19 related malicious misinformation (BBC, 2020b). Proportionally much of the content originated from peripheral far-right groups. The malicious misinformation centered around claims that Covid19 was artificially created with malicious intent, targeting Judaism (claiming a Jewish global elite created Covid19), Islam (Indian posts claiming the virus was of Muslim origin), global elites (conspiracy theories claiming Bill Gates ordered the creation of the virus to profit from its vaccine), and other misinformation encouraging distrust of national authorities. As the BBC pointed out, it was not the number of posts that was most concerning, but the fact that they managed to become part of mainstream FB content (2020b). Thus, harmful content was normalized into being information on Covid19.

Regarding the specific posts and groups which the BBC had located, they received the reply from FB that:

> We have removed a number of links […] for violating our policies on hate speech and the spread of harmful information. When a post does not violate our policies but is deemed to be false, we reduce its distribution and show warning labels, which for 95% of the time means people do not go on to view the content" (BBC, 2020b).

So, content that FB deems false that is not hateful or harmful remains on FB's platform. But these posts are labeled "false", effectively stopping most users from accessing the content. Furthermore, its visibility on the platform is artificially reduced. What FB considers harmful content is taken down.

In section 3.1 I already pointed out that FB self-regulates user content based on self-determined conceptions of what is false and harmful. To the degree openness to dialogue is possible on FB, FB moderates the conversation and inserts its self-regulated opinions on what is harmful or untrue.

Discussion

The traditional account of FE stresses the importance of honest dialogue for democratic legitimacy and a progressive understanding of truth in society. FE results in and protects the expression of a diversity of viewpoints. Not only do Mill and Dworkin argue for the freedom to express one's side of an argument, citizens need to hear the other side of an argument too. To this end, people need unmoderated symmetrical access to each other's (harmless) expressions.

Online platforms are not directly suited for hearing the other side of an argument. The business model of social media platforms like FB are built to optimize the collection of personal data by creating an individual user experience optimally aligned with their preferences. The content one meets reflects one's interests and social relationships. The only factor giving a FB user access to a diversity of opinions is if one's social network reflects such diversity. This can be argued to mean that a user has access to a diversity of opinions if this is a social preference one has established in one's circle of FB friends.

But, even if a user has a spread of FB friends, representative of societal diversity, the content accessible is self-regulated by FB based on self-determined conceptions of what is false and harmful. On a Millian account, FB is not justified in moderating or "tagging" presumed falsehood. Falsehood is accounted for in the progressive dynamic dialogue creates.

The confirmative working of an online echo chamber is on a Millian view the negation of potential dialogue. Even the truest opinion, according to Mill, when going unquestioned and

unchallenged over time, turns into a stale and empty dogma (Mill, 1989, p.37). This is the opposite of the lively transformative dynamic of an experienced and self-won truth.

Since physical distance plays no role online, and if language is not a barrier, online platforms increase opportunities for community formation based on commonalities other than geographic location. Zuckerberg confirms that people most often create online communities around a shared view or interest (2019b). But FB adds an enormous quantity of voices to one's network, possibly from a diversity of geographic locations. This can add a type of diversity to expressions a user has access to that was not physically possible before social media platforms. But it seems plausible that Mill is more concerned with experience with a diversity of viewpoints concerning one's national society, than with being confirmed in one's views, internationally. And since democratic processes a user partakes in concern the nation one is a citizen of, Mill's argument for connection across one's own society's diversity seems to have more weight for FE than FB's potential for a large quantity of global connection based on commonalities.

FB's use of an algorithm provides people with echo chambers of more of the views they already endorse and less of those the algorithm predicts they oppose. But Zuckerberg tries to oppose this image of FB as an individualized echo chamber:

> Research from the Reuters Institute also shows people who get their news online actually have a much more diverse media diet than people who don't, and they're exposed to a broader range of viewpoints. This is because most people watch only a couple of cable news stations or read only a couple of newspapers, but even if most of your friends online have similar views, you usually have some that are different, and you get exposed to different perspectives through them (Zuckerberg, 2019b).

So, Reuters' research shows that online users tend to use more sources than offline persons, thereby accessing a more diverse diet of information than people who read their one newspaper over morning coffee. But, Zuckerberg's attempt to include FB in this tendency based on one's FB-friends with different views seems rather thin. It puts a lot of democratic responsibility on the fact that the online FB-content that will always show up on your feed, regardless of the algorithm's predictions about you, are expressions posted by FB-friends.

Zuckerberg claims FB has a duty to support a diversity of ideas. But this diversity is reflected in the totality of content the platform hosts. This totality is not the content a user is

naturally presented with. Therefore, social media like FB do not support dialogue across a natural diversity of views.

FB may object though, that Mill says nothing about a natural or easy access to a diversity of opinions. Maybe it is the duty of the individual user to actively gain access to such diversity and dialogue. Then all FB needs to do is not hinder them in doing so. FB's duty in line with FE would then be to not interfere with the free access to information and dialogue providing connections. This is a freedom FB supports through noninterference to a degree. But there are three important factors reducing this noninterference.

Firstly, FB's algorithm distorts what would be a natural presentation of the expressions existing online. Secondly, FB's content moderation practices mean not all content is allowed on their platform, therefore the diversity of content accessible on FB is a censored sum without democratic accountability for why content is excluded. Thirdly, FB has the capacity to change user access to content and users at any time, a user has no claimable rights. Therefore, social media with content moderation policies, algorithms, and full power over the content on their platform, cannot be said to realize noninterference with access to content in line with the HRFE.

Of course, if FB would allow a natural and symmetrical access to a diversity of viewpoints this does not naturally create dialogue. But, the active limitations on such access do make establishing dialogue unnecessarily hard. In addition, it is FB's processes of content moderation that encourage a user to indiscriminately express opinions rather than to look for informed discussion with an honest intent. Thus, FB's structure and policies negatively impact opportunity and motivation for discussion between users with challenging views with a mutual intent for honesty and potentially human growth and a progressive understanding of truth.

On Mill's view of a perfectionist progress in both individuals and society in which honesty plays an important role as a duty, FE is not meant to be a sheepish toleration of all meaningless and hurtful expression people are capable of. So how does this progressive vision hold up against the flood of unkind memes, willful misinformation, and malicious hoaxes that social media host? Mill assumes the honesty of the speaker. How does this relate to the online deliberate propagation of fake news and "bullshit" as Frankfurt termed it (Frankfurt, 2005)?

Mill shows a fundamental trust in a naturally dynamic process that individuals and therefore society undergo by sincere participation in dialogue. Thus, all harmless content on social media can feed into society's progress. According to Mill, then, social media should encourage, or at least not hinder, the potential of honest discussion across a diversity of opinions. Rather than maintaining such an expectation of sincere user attitudes, FB maintains a lowest possible

threshold for publishing opinions and boosts the virality of the most popular ones, regardless of honest intent, informed content, or the (lack of) value of an expression for public discourse.

But Mill is not naive about the majority of people not being interested in striving for progress or honesty. Mill's perfectionist ethics seem to add to the idealistic vision of his ethics, but he acknowledges that these ethics don't attract much active practitioners. For the majority, then, it is the less idealistic traditional account of FE, meaning safeguarding public discourse from the voices that intent harm, that counts. Mill argues for the greatest possible freedom in public discourse otherwise.

Partially in line with the traditional account of FE, FB regulation reduces the presence of what FB deems to be harmful expression. Therefore, FB seems somewhat aligned with the Millian minimum condition of enforcing the harm principle, though FB's moderation of harm lacks democratic legitimacy. But FB does not encourage its users to strive for moral duties such as challenging dialogue, honest intent, or progress towards a greater or more lively understanding of truth.

3.4 Real and morally accountable people

The traditional FE accounts for expressions belonging to real people with real names. According to Mill, moral (and presumably also legal) accountability is an important aspect to expressing one's opinion. Accountability requires for a speaker to be identifiable by the social environment one is accountable toward or held accountable by.

In this section I discuss how FB's user policy requires users to be identifiable as a real person with a real name. This is to increase trust on the platform and to prevent harm (Zuckerberg, 2019b). Even though FB requires users to use their real identity on its network, the platform hosts a great amount of what I call "anonymous accounts". By anonymous accounts I mean those accounts that do not represent or identify one real user. Some of these accounts are used to cause harm, both online and offline. Other anonymous accounts, however, can be argued to increase rather than threaten the HRFE.

Anonymous accounts

According to Zuckerberg, FB has had most success preventing harm through ensuring speaker authenticity (Zuckerberg, 2019b). Ensuring speaker authenticity includes removing accounts that are fake. Many anonymous accounts are run by AI and are often run in large cooperating clusters. Others are run by a person under a false identity or tie several fake

accounts to one person or organization. FB's terms of service require speaker authenticity (FB, 2020). The requirement does not differentiate between users having a false identity with the purpose to deceive, and those who have what qualifies as a false identity as an expression of an alter-ego, artist name, identity choice, or other arguably harmless purposes.

As Zuckerberg explains, "Focusing on authenticity and verifying accounts is a much better solution than an ever-expanding definition of what speech is harmful" (2019b). Pinpointing accounts with qualitatively and quantitatively major reach, Zuckerberg says that FB "now require[s] you to provide a government ID and prove your location if you want to run political ads or a large page. You can still say controversial things, but you have to stand behind them with your real identity and face accountability" (Zuckerberg, 2019b). Identity is thus a condition for having a voice on FB.

Anonymous accounts can be used to increase the impact of one's voice online (for example, one person creating multiple anonymous accounts to promote a cause through a quantitative increase in response). But it is especially anonymous accounts in large coordinated quantities that have an extreme capacity to manipulate and cause real harm, online and offline. An example is the systematically spread anti Rohingya propaganda on FB by Myanmar's military (Mozur, 2018). For about five years, the military used fake accounts to anonymously spread malicious misinformation about the country's Rohingya minority. Thus, the military purposely abused FB's extensive user reach to spread hate targeting one of Myanmar's minority groups. This online propaganda campaign finally exploded in the physical world, resulting in real life murders, rape and the forced mass migration of many of the Rohingyas (Mozur, 2018).

As mentioned before, FB user communities around shared views and FB's content moderation that reinforces this tendency, create fertile ground for harmful manipulation. The degree of psychological manipulation on other users becomes exponentially greater when it comes from an orchestrated group of what seem to be peers within an online community around a shared view (Donath, 2019). For example, when a regular user meets one counter opinion posted by an AI run fake account, the manipulative or deceptive effect may be minimal. But when three or four or twenty different coordinated (fake) users agree with each other on this counter opinion, a user may be swayed by the peer pressure that comes from this dynamic. When discussing matters of importance socially, politically, psychologically or economically, experience has shown the real harm and danger involved.

Harmless intent

User anonymity online can have a harmless or harmful intent. Anonymity can be used to express oneself unbounded without facing consequences (for example, net trolls anonymously posting hateful comments or threats (Gorman, 2019). But user anonymity may also be a partially misleading label for an altogether different phenomenon. There are FB users who are trying out a different identity. To FB, since the user does not use the name they use in everyday life, this is an anonymous- or fake account. But these users don't necessarily intent anonymity, they intent to be a different aspect of themselves. For example, a closet gay person who uses a fake account to participate in the online gay community (Grinberg, 2014).

Anonymity, or its equivalent pseudonyms, have made possible many goods. A name can generate unwanted attention to general prejudice or misconception. There are many examples of contributors to political resistance or artists who preferred the anonymity of a false name to avoid either unjustified negative attention or real physical harm.[24] In many of these cases, it can be argued that their FE was protected or enhanced by anonymity.

Judith Donath (1962) is a Professor of media arts & sciences who does research on how new technologies transform identity and social reality. As Donath describes it, "[i]dentity experimentation – from attempting to pass as the opposite gender to creating an alter persona to enjoying the freedom of speaking from behind an anonymous screen – is for many a goal in itself" (1995, p.7). This phenomenon is not unknown in physical society, just a lot harder to achieve with a physical body. As Donath points out, "[w]hile in the real world, gender deception takes a great deal of effort, in the online world it is simple: use a name typical of the opposite sex" (1995, p.7).

It may be that the relative ease of gender deception online lowers the boundary for using it for other purposes than identity experimentation. The ease of abuse could legitimize an altogether prohibition of the phenomenon. But if identity experimentation is a significant psychological need for some people, anonymous accounts may be argued to go in under FE.

How does FB argue for its conditions for user identity and what are they?

[24] Many artists choose a more English name to avoid unwanted distractions connected with their nationality. Freddie Mercury for example, was called Farrock Bulsara. Author J.K. Rowling published Harry Potter first without revealing her gender since the publisher thought the target audience of young male readers would not want to read a female author's books. Jane Austen wrote under the name "A Lady". People active in political resistance use pseudonyms for reasons of safety. Nelson Mandela worked in the underground resistance against South African apartheid as "the black pimpernel".

FB's user conditions for identity

Under section 3 of their legal terms of service, called "your commitments to Facebook and our community", FB specifies its conditions for FB users (Facebook, 2020):

> 1. Who can use Facebook
>
> When people stand behind their opinions and actions, our community is safer and more accountable. For this reason, you must:
>
> - Use the same name that you use in everyday life.
> - Provide accurate information about yourself.
> - Create only one account (your own) and use your timeline for personal purposes.
> - […].
>
> We try to make Facebook broadly available to everyone, but you cannot use Facebook if:
>
> - You are under 13 years old.
> - You are a convicted sex offender.
> - We've previously disabled your account for breaches of our Terms or Policies.
> - You are prohibited from receiving our products, services, or software under applicable laws (Facebook, 2020).

FB's reasons for assuring speaker authenticity are based on a wish to secure user accountability and the safety of the community. It seems likely that by "community" FB means its own online community of users, since it is phrased "our community". But, since FB in its conception of harm includes effects on real people in offline situations, logically society concerns both FB's online society and those parts or aspects of offline society affected by FB's content and users.

The reason FB gives for their user conditions seems aligned with conditions for offline expression. Accountability and security in relation to the community are values compatible with Mill's values and harm principle. Furthermore, the terms of use come down to the fact that users should be one person, one voice, identifiable under the same name and facts that make them identifiable in everyday life offline. These too are values in alignment with Mill's view that expression and opinion should be tied to real and morally accountable people.

But when considering FB's conditions for denying a user access, the regulation seems arbitrary and grounded in other concerns than those stated earlier in the section. For example, children under the age of 13 are not allowed to join FB. This is because in the USA, the *Children's Online Privacy Protection Act 1998* (COPPA)(USA) prohibits apps and websites

from collecting personal data from children under the age of 13. This law is irrelevant for other parts of the world, since similar foreign regulations set different age limits. For example, in 2018, the European Union passed a law called the *General Data Protection Regulation* (2018) that determines that all children under 16 in the EU need parental consent for user accounts on social media sites and the collecting of personal data online. So legally, the age limit is not the same for all user nationalities.

Furthermore, FB's determination of a user's required minimum age, seems unrelated to the argument for accountability and safety FB claims to base its terms of service on. If speaker authenticity ensures user accountability and the safety of the (online and offline) community, then a user's minimum age should logically correlate with legal accountability. Instead of choosing accountability and safety, FB prioritizes the possibility of personal data collection. This makes the opening statement concerning the reasons behind who can use FB less credible.

A similar objection to arbitrariness concerns FB's excluding of convicted sex offenders. I see two objections against this user condition. Firstly, FB does not account for why this group of convicted offenders are different from all others. Why not, for example, also exclude other offenders like convicted murderers or those who have committed large scale fraud? Secondly, it is unclear whether FB differentiates between convicted (sex) offenders who served their sentence and those who have not. If FB does not make such a differentiation, the question is on what grounds FB chooses to hold people accountable for a crime for which legal systems consider their legal accountability as completed.

Lastly, making compliance with FB's self-regulated "Terms and Policies" a condition for FB users, gives FB a free pass on restricting access to its platform without democratic accountability for their reasons.

So, while the reasons given for FB's user conditions seem to support the traditional condition of real and morally accountable speakers, the scope FB determines through its restrictions seems arbitrary and unrelated to the original reasons.

Discussion

We need new ways of thinking about identity online. Recognizing the need for ways of assessing identity and its potential online, Donath proposes the use of theoretical models from other disciplines (Donath, 1995). In physical society, identity, especially when considering it in relation to representing one voice in a democratic society, is most commonly associated with a locatable physical body. As Donath says, "[i]n the physical world there is an inherent unity to the self. The body, problematic as its philosophical relationship to the self may be, provides a

compelling and convenient definition of identity. The norm is: one body, one identity" (1995, p.1).

But the non-physicality of the online world refers to different aspects of reality than the anchor a body provides. As Donath points out, virtual reality is about information more than it is about matter. And information obeys different laws than matter does, it "spreads and diffuses. […] there is no law of the conservation of information. The inhabitants of this impalpable space are also diffuse; […]. One can have as many electronic personas as one has time and energy to create" (Donath, 1995, p.1). But there is a physical body sitting at the computer screen that grounds the diffuse online identity. The offline world and the online world are connected, it is just that the relation between the two can be unclear or misleading.

Donath, in discussing ethics in relation to artificial entities, addresses the human tendency to ascribe some form of personhood or sentience to AI we interact with (Donath, 2019). Humans do this even when they know the voice represents AI, for example in the case of robots performing basic care tasks (Donath, 2019). But this tendency for ascribing personhood and its effects are much greater (with an enormous potential for harm) when people are unaware of the fact that they are interacting with AI.

Anonymity can be considered a value or a threat. As a value, it gives an opportunity of identity experimentation that physical reality does not allow for with the same ease. But anonymity gives equally greater capacity for harmful conduct, a reduced sense of accountability, and an erosion of community (Donath, 1995, p.2).

Would it be possible to apply the harm principle to the question of identity, seeing as an aspect of FE? The Millian view is that anonymity reduces individual responsibility and makes impossible social accountability. It seems obvious that anonymity makes it harder to hold the individual accountable for harm incurred. But that does not mean that anonymity is harmful to others. It may be a positive contribution to liberty overall. So, to claim anonymity is a wrong because it may be used for a wrong seems too strong a conclusion.

Deception brings gains to the one deceiving and losses to the deceived party (Donath, 1995, p.2). When communicating with an unfamiliar person there is much to be gained from discovering deception before experiencing its detrimental consequences. Often we can find out through communication. Expressions can either reveal real knowledge or personality traits, or just reflect social convention or empty claims (Donath, 1995, p.3). For example, when talking about divorce, something said may reveal the presence or lack of insight only the experience can provide. And an empathic sounding response may be real empathy or just a collection of the right sounding words.

The extent of harm deception causes depends on the context and the kind of information shared. As Donath points out, deceptive misinformation online can be very detrimental to its audience if it concerns a false but legitimate seeming medical doctor giving injurious medical advice (1995, p.5). But users tend to learn from experience. So, if a similar scam happens often, the tendency to believe online medical statements is reduced enormously. Consequently, it is the legitimate medical expert giving legitimate medical advice who suffers a loss of credibility (Donath, 1995, p.3). This last consequence is an example of what FB likely means when it prohibits false accounts altogether on the grounds of its erosive effect on trust.

Should FE include the right to anonymous expression? Mill disagrees, preferring even the ballot to be public to increase responsible behavior (Ten, 1998, p. 374). The correlation between being held accountable by one's peers or society and a person's felt need to act responsibly makes sense according to moral psychology. Moral theory praises the morality of a person who acts as responsibly without witnesses as when in public, but moral psychology acknowledges that this is a final stage of moral development reached by few. Kohlberg's theory on progressive stages of moral reasoning suggests that the more common moral capacity is for people to need norms with external accountability to support them in being responsible (Kohlberg, Levine & Hewer, 1983). Since Mill pointed out a similar concern when stating the ballot should not be secret, I will treat Mill's account of FE as having this view of moral psychology as one of its premises. Anonymity as an aspect of FE raises concerns of lack of accountability. The abuse of anonymity on FB in the form of for example harassment by online trolls illustrates this concern (Gorman, 2019).

But what if user anonymity protects the user from harm? Anonymity may for example protect a vulnerable person from bullying. As discussed in the section on Mill, though Mill does not think psychological harm should be included in the harm principle, Go points out that our current knowledge of psychological harm gives us knowledge and therefore reasons Mill lacked. Therefore, online bullying has a strong argument for being included in the conception of harm.

But FE means that the government should protect citizens from harm enacted by others. Being secure from harm is a claim a citizen has against its government. Anonymity is not supposed to take the place of this right as an alternative solution. Therefore, a user's vulnerability is not an argument for the legitimate use of anonymity, at least not offline. But who is the protecting government online? The core argument for identifiable voices is that it makes an individual accountable. But what if it is not safe to be accountable online, can accountability still be a legitimate FB user condition?

Accountability is relational. Offline, a person expressing her opinion is accountable to the liberal democratic society she is a citizen of. Online, accountability for expressed content is legally only to the offline society the user that wrote the content is a citizen of. Online, FB can remove a user for breach of user conditions but cannot enforce legal implications. Thus, FB does not have the capacity to protect a user from real harm in the real world. Logically, a FB user is not accountable to FB for anything other than *reasonable* user conditions. It seems reasonable to say that a user's safety concerns in the real world should trump FB user conditions. So, whereas it seems offline society has legitimate reason to claim civil accountability because it provides the duty of ensuring the correlating safety, it seems that FB's accountability claim is unreasonable without correlating duties.

Another concern is that offline government cannot provide the duty of online protection of its citizens to the degree it should either. Certain legal claims concerning online content function well to the degree they are similar to offline expressions. But, as the British House of Commons concluded after its investigation into detrimental online interference with the Brexit election, "British election laws were not fit for purpose" when it came to the violation of democratic processes through online campaigning on FB (House of Commons, 2019). If a user sharing content online is not sufficiently protected from harm by neither FB nor offline government and legal systems of the country one is a citizen of, online anonymity seems to be a rational alternative for self-protection.

3.5 A closed, physical society

The traditional account of FE is implicitly situated in a physical and closed society, undisturbed by influences outside itself. The Millian account presented in "On Liberty" focuses on the political relations between a government and its citizens (Mill, 1989). This account may run into problems when being applied to the globalization of both political relations and expressions. Contemporary offline society is internationally related and affected by social media.

Social media like FB are self-regulating, independent from liberal democracy, yet have an undeniable impact on offline democracies. FB and online society represent an even stronger degree of global connectedness than contemporary liberal democratic society and is defined by information rather than physicality. FB is, except as a physical business located in the USA, not

accountable to liberal democracies because the law is not caught up with online reality yet.[25] In the freedom this creates for FB, the company has established its own policies and guidelines for the regulation of expression.

FB is not accountable to liberal democracies

FB is a tech company providing an online platform for expression. Being a privately owned and therefore self-regulated company, FB does not need to comply with the HRFE and its role as a protection of democracy. My stance is that, because of FB's power over the online speech of such a large part of the global population and its impact on liberal democracies, FB should be accountable to liberal democracies. Then, liberal democracy could extend its FE to online expression.

FB self-regulates its own principles and policies to abide by. Zuckerberg is careful with the legal implications of wording. He mentions FB is "inspired by the First Amendment" (Zuckerberg, 2019b). And, "we look to international human rights standards […]" (Facebook, 2019b). Being inspired by a principle or law does not mean one strives to abide by it. FB could choose to develop its policies in alignment with the HRFE but does not do so.

On FB, users are accountable for their expressions to the nation they are citizens of. FB has no legal accountability for content on its platform. This is made legal by §230 of the *Communications Decency Act 1996* (USA), which is meant to protect online platforms from endless liability claims against them regarding content created by their users (Bowers & Zittrain, 2020, p.2). But in addition, it means that FB's content moderation policies practices have no legal implications for FB as a platform provider, but fall under FB's own FE.

Legal scholar Benjamin F. Jackson discusses FB's handling of censorship and FE and the legal problems that come up (Jackson, 2014). Jackson points out that, since current law does not adequately relate to online content and its platform providers, there is no good way of showing the relevance of laws implementing the HRFE in relation to online content. This in turn means there are no cases creating legal precedence that could determine the law's applicability for future relevance (Jackson, 2014). So, unless new and specifically relevant legal solutions are created, the current legal framework is stuck in a loop of impotence regarding online expression.

[25] As David Carroll, an American professor in Media Design, says concerning FB, "this is a company that is a super-state. And the only nation that has jurisdiction over it is ours" (Amer & Noujaim, 2019). FB pays taxes in the USA.

Rather than recognizing the importance of finding a way for tech companies to bridge online expression with offline liberal society, Zuckerberg focuses on further developing the structural changes social media have brought about. He claims that the reality of individual expression online has introduced a new power structure. He calls it

> a Fifth Estate alongside the other power structures of society. People no longer have to rely on traditional gatekeepers in politics or media to make their voices heard, and that has important consequences. I understand the concerns about how tech platforms have centralized power, but I actually believe the much bigger story is how much these platforms have decentralized power by putting it directly into people's hands (Zuckerberg, 2019b).

Zuckerberg claims that online platforms like FB have given people a direct access to a power that bypasses traditional forms of power. This had been true if FB had been true to the original intention for the internet to be an unmoderated free society. But Zuckerberg does not seem to appreciate the fact that FB moderates these expressions without giving users power over how, when, and to what degree FB does this. Since FB does not provide users with such (preferably democratic) decision power, factually, it is tech companies like FB that constitute the new power structure Zuckerberg refers to, not the people.

Zuckerberg implicitly acknowledges difficulties related to the self-regulatory power of tech companies online, when he expresses his concerns over a Chinese take-over of online power. Identifying himself with a predominantly American value-system, he says,

> [t]his question of which nation's values will determine what [online] speech is allowed for decades to come really puts into perspective our debates about the content issues of the day. […] If another nation's platforms set the rules, our discourse will be defined by a completely different set of values (Zuckerberg, 2019b).

Thus, tech companies' self-regulated values determine values and policies (international) users abide by on these companies' online platforms. And it is tech companies that define the rules for online discourse.

Being an American company with global reach, with no political ties or legally defined duties towards its users, from what or who does FB derive its policy decisions? When discussing FB policy regarding content moderation, Zuckerberg rarely argues from a principle but tends to take as his reference social trends or tensions, such as "shifting cultural sensitivities and

diverging views on what people consider dangerous content", though he claims these should not determine FB policies on expression (Zuckerberg, 2019b).

Zuckerberg's stance on social trends and cultural sensitivities is reminiscent of Mill's conception of morality as a time-dependent social phenomenon. As opposed to ethics, which is built on a (timeless) principle, morality is in Mill's view mostly a product of the current preferences or resistances of society as a collective. Morality is thus very much an expression of a time-spirit (Mill, 1989, p.10-1).

Having a focus mainly on morality instead of ethics could explain the seeming arbitrariness of some of FB's policy decisions, especially the odd conglomerates of restrictions. But in Zuckerberg's defense, online expression is a rapidly changing, relatively new, and impossible to foresee phenomenon. When a new way of incurring harm online arises, its reach is enormous almost instantly. FB needs speed and flexibility in dealing with issues as they arise. Still, a principle can be argued to add to the speed, effectiveness, and overall meaningfulness of such dealings.

Zuckerberg also considers the American legal framework for free speech (i.e. the first amendment):

> Some people argue internet platforms should allow all expression protected by the First Amendment, even though the First Amendment explicitly doesn't apply to companies. I'm proud that our values at Facebook are inspired by the American tradition, which is more supportive of free expression than anywhere else. But even American tradition recognizes that some speech infringes on others' rights. And still, a strict First Amendment standard might require us to allow terrorist propaganda, bullying young people and more that almost everyone agrees we should stop — and I certainly do — as well as content like pornography that would make people uncomfortable using our platforms (Zuckerberg, 2019b).

Besides noting that companies are not legally bound by the first amendment, Zuckerberg claims it would be too free as a framework for FB as a platform to use for regulating expressions. It is neither the public (social trends and -tensions) nor legal frameworks that determine FB's regulation of online expression. Policy and regulation happen according to premises and processes known only to FB itself.

The independent oversight board for content moderation

Zuckerberg acknowledges the problematic concentration of FB's power over its users' expressions (2019b). A natural solution would be to harmonize FB in a meaningful way with existing democratic structure and institutions. But instead of moving towards the liberal democratic structures that govern offline society, Zuckerberg strengthens FB's independence by introducing a new extension of FB's current business structure, the Oversight Board:

> That's why we're establishing an independent Oversight Board for people to appeal our content decisions. The board will have the power to make final binding decisions about whether content stays up or comes down on our services — decisions that our team and I can't overturn. We're going to appoint members to this board who have a diversity of views and backgrounds, but who each hold free expression as their paramount value (Zuckerberg, 2019b).

FB establishes an oversight board that users can appeal to when they disagree with FB's decision on a specific incident of content moderation.

The *Oversight Board Charter* launches the framework for what Zuckerberg called an "experiment in independent governance on expression" (Zuckerberg, 2019a). The charter establishes the oversight board is an extension of FB, adhering to FB's community standards. The purpose of the oversight board is "to protect free expression by making principled, independent decisions about important pieces of content and by issuing policy advisory opinions on Facebook's content policies" (Facebook, 2019c, p.2). By establishing an organ for appeal, FB seeks to increase the transparency and legitimacy of their content moderation of user content. Appeal decisions are in line with FB's community standards.

The possibility of appeal is FB's first opening for a democratic process. But there are several factors that reveal just how minimal this opening is.

Firstly, content appeals are within the framework of FB's self-determined policies. FB' community standards and content moderation policies are not attained through a democratic process. nor do they adhere to legal- or political frameworks outside FB (Facebook, 2019c).

Secondly, the oversight board strengthens FB's independent business structure. FB's power to affect their user's online expression without legal accountability for these policies is maintained.

Thirdly, the only power the oversight board has, is to make binding decisions on individual cases of user complaints on content moderation, in line with FB's community standards. Other

functions are limited to the giving of non-binding advice to FB. The oversight board as an appeal body does not have the power to change FB policy.

Fourthly, none of the bodies involved (i.e. the oversight board, the trust, FB) are elected through a democratic process, but rather through FB's internal self-confirming processes.[26] From considerations of democracy and external accountability, the power structure proposed in the charter seems to confirm FB's position rather than diminish it.

Discussion

Mill's closed and physical society does not account for online expressions on self-regulated platforms, independent from offline liberal democracy yet affecting its citizens' FE and its democratic processes. Due to the enormous quantity of expressions on FB, it seems a threat to the HRFE's function and principle as a right, that this content is not protected by the HRFE. Nor is the content on FB's online platform currently clearly accounted for within each specific liberal democracy. Thus, online expressions are outside of the legal scope of either a nationally or a globally focused HRFE. This may seem a practical legal concern, but it is also a philosophical one. The relevance for this discussion is, that the lack of legal power to include social media in current national legal systems, means that content on FB, to a large degree, effectively exists outside of any liberal democracy and its account of the HRFE. But FB content has come to affect public debate, both offline and online, and therefore the function of FE as a protection of liberal democracy.

In liberal democracies it is FB who has the most restrictive effect on its users' HRFE. Liberal democracies practice a much greater restraint towards restricting the scope of expression. Nor do they affect expressions without the possibility of democratic influence on these policies in the way FB systematically does.

But does FB really impact liberal democracy that much? Yes. Here are two arguments that support this claim.

Firstly, FB rules the online market - most social media platforms are American, and most American platforms are owned by FB Inc. When it comes to online social networks, FB has no real competitors (CBSN, 2018). Social media are attractive to users to the degree one's real or

[26] FB appoints the initial members of the oversight board. Afterwards, Facebook and members of the public may "propose candidates" to the oversight board (Facebook, 2019c, p.4). After the initial establishment of the oversight board, the members of the oversight board select new members, but the formal appointment and the possibility of elimination lies with the trust (Facebook, 2019c, p.4). The trust is another FB organ, members of which are always appointed by FB. Instead of implementing a democratic or otherwise public process, FB seeks to legitimize the independence of the oversight board through the "middle body" of the trust (Zuckerberg, 2019a).

preferred or potential social network is present on the platform. FB is the largest online platform and is therefore the most attractive to users.

Secondly, as Dan Patterson of "Techrepublic" points out, "Facebook is an unelected hegemonic power that is dominant in our everyday life, whether you use the platform or not", we only have to consider the rigged American presidential elections of 2016 to see the truth of this (CBSN, 2018). Though the voters manipulated to vote for Trump were FB users, the consequences are real for all involved offline.

In contemporary society, globalization of expression and the many forms of external influence that are present are at least as urgent as matters of national importance. Did the Millian account become outdated with the advent of globalized communication platforms?

Zuckerberg believes the internet to be new because of the speed and reach of online expressions, "[i]t's empowering that anyone can start a fundraiser, share an idea, build a business, or create a movement that can grow quickly" (Zuckerberg, 2019b). Noting the harmful side of this phenomenon, it is hard to estimate the reach and consequences of events such as when Russia's IRA interfered in the 2016 presidential elections in the USA (Zuckerberg, 2019b). Another related phenomenon is virality of harmful content, like harmful misinformation or fake news. All of these online phenomena have the potential to affect offline reality.

Internet seems to reverse the Millian corn-dealer case into its exact opposite. If an anti-corn-dealer meme goes viral, FB facilitates the organization of seasonal workers that feel victimized. The fact that FB aims at community building has been abused to create groups joined together in the fight against a perpetrator or oppressive force, online and offline. Social media like FB have transformed the Black Lives Matter movement, originating in the USA, from a minority movement to a viral online force that organizes real-life protests for different related causes globally (for example in Australia, Canada, UK). Online movements around common enemies or accusations have shown they can cause physical harm in the real world. But this reversal of the Millian example does not mean Mill's harm principle is inadequate. It means online expressions have altered what we should recognize as a context that can cause (physical) harm.

Just like Mill, Zuckerberg argues for regulating speech to protect from imminent physical harm. Mill's conception of media and its effect on society and individuals may fail to account for the enormous impact social media have. Mill's account remains applicable in its normative essence, but liberal democratic society needs to find a way to include online expression on social media like FB in the legal reach of its FE.

3.6 Summary

The fact that FB is not regulated into complying with FE does not mean it does not act in line with FE, FB may argue. FB claims it is an instrument of FE. The company's examples of this are that it's a free platform making possible having an online voice for everyone with an internet connection. Furthermore, the company facilitates connection, community building around common interests, reduces advertisement disadvantages for small companies, and chooses to act on certain kinds of harm online.

In this chapter I have argued and given examples of why FB's commercial image as an innovative force for online FE does not match the conditions that liberal democracies traditionally have for freedom of expression. The crucial discrepancy between the traditional account of FE and online expression on FB originates from, I have suggested, FB's self-regulating power.

Mill points out the threat of a government's assumption of infallibility about its conception of truth. He defines the assumption of infallibility as a government claiming it has the right and (sufficiently) complete understanding of what is true and right, and then decides that question for others. FB claims it has a right and sufficiently understanding of what is true and right, furthermore, imposes this on its users. Can a tech company be accused of acting on an assumption of infallibility? FB is not a state, so there is no legal reason to expect FB to provide any of the Millian goods unless they themselves promise they will.

This chapter has given an insight into how FB's policies are not in alignment with the traditional account of FE, which is how liberal philosophers have conceptualized the HRFE. As such, expressions on FB do not contribute to the role that FE plays in liberal democracy and can be argued to harm it. If we accept this first conclusion, then my thesis can inform society on what I see as the problematic absence of respect towards the HRFE on social media like FB.

The reason I think the absence of the HRFE on FB is problematic is because of the power and reach social media have.[27] Zuckerberg claims the early FB platform had many positive effects on university campus. It connected people, contributed to entrepreneurial impulses within the campus community, and inspired political engagement with campus regulations. Based on Zuckerberg's example, it seems FB worked admirably at that scale and while embedded within the larger framework of campus life and regulations. But now FB is no longer

[27] The quantity of users, the global reach, and the enormous regulative power social media like FB have over its users' expressions make social media into a power similar to (other) transnational companies. The power of transnationals is in certain areas greater than that of nations.

embedded in that manner. It hosts a global network of people spread out over many different nations. And the national legal frameworks are lagging behind in how to deal with the new challenges social media like FB pose. Much of the power to establish and regulate this new situation lies with tech companies themselves.

In the next chapter I will discuss if online expression on social media like FB should be regulated by liberal democracies.

4 Should expression on social media be regulated by liberal democracies?

My conclusion from chapter 3 is that the way social media like FB regulate online expression does not further liberal democracy, but rather undermines it. The dynamic that is revealed, is that physical liberal democratic society has limited possibilities of affecting expressions on social media and the policies moderating them. But expressions on social media have a real impact on liberal democracies offline. Cambridge Analytica influencing elections worldwide is no minor example of this. And the role of FB in spreading hateful propaganda for years leading up to the Rohingya massacre, means FB played some part in what turned into an inhumane horror for people's physical existence. Also, it will never be clear how often the Black Lives Matter campaign included fake protests organized by Russian actors (Amer & Noujaim, 2019).

Assuming social media like FB will develop and expand rather than disappear, what does this mean for our conception of FE? Given the impact of online expression on contemporary society, I think FE should be respected on social media.

But having established that there are large discrepancies between traditional FE and online expression on social media, how do we harmonize between the two? In what direction do we harmonize between online and offline expression? If we want to keep liberal democracies as they are, then the most logical solution is that we should regulate social media into adherence with the HRFE. But this solution likely faces problems.

In this chapter I briefly investigate different ways of approaching the challenging discrepancies between online expression on social media and traditional freedom of expression. I have already argued extensively that, considering the function of traditional FE, social media should not remain as they are.

Still, first, I briefly consider possibilities of adjusting FE in its legal form of the HRFE to current social media reality. I discuss why I find these scenarios problematic.

Next, I consider if we should impose government regulation on social media. I cannot provide a definite answer, but I sketch some ideas and problems these may face.

What if we rethink the HRFE to adjust to social media?

Rethinking the HRFE means, first, assessing whether the right still achieves what it is intended to achieve, specifically in relation to its role as a protector of liberal democracy. Then considering whether or what changes are necessary to achieve the aim of the HRFE.

Much of both private and public discourse happens on social media. For liberal democracy to function well, every individual in society should be free to hold their own opinions and express these without interference if doing so does not harm someone else. Since social media have created a fundamental change in the possibilities we have for expression, some may argue that rethinking the HRFE may be appropriate.

FB might argue that adjusting the HRFE to the reality of its platform would enable us to keep the benefits online expression on social media provides. What are the goods that come from social media like FB? I discuss two main goods FB might suggest.

Firstly, tech companies' capacity for innovation of online expression. FB has been at the forefront of online innovation, turning social media into global social networking and business platforms. It is likely that FB's innovative power and speed is dependent on its self-regulation. Self-regulation allows the business to initiate new technologies and adjust policies, not held back by anything but its own business and image considerations.

But the cost of self-regulated innovation is that offline society loses its right to democratic deliberation on changes affecting online FE and physical society. For example, with technological innovations follow ethical concerns. Not everything technology makes us capable of doing, is something society wishes to make a possibility. FB's focus on innovation has created situations in which the company cleans up after the fact, with limited success. Livestreaming for example, has given FB users and entrepreneurs a new opportunity for sharing content. But, as previously mentioned, the same tool has been used for the instant and global broadcasting of extreme physical harm. FB continues the use of this tool despite its continued abuse and the impossibility of sufficient regulation to prevent the live broadcasting of harm.

The second good FB might argue for, is its facilitation of an online voice for all who have an internet connection. FB's self-regulated services come with the promise of never requiring a monetary user fee. So, having an internet connection is the only challenge a FB user faces, and this makes FB's services very accessible.

But the current form the transaction between FB Inc. and a FB user takes entails three consequences that are not obvious to all users. Firstly, all user expressions are archived by FB. When a user publicly or privately publishes content anywhere on FB, she hands some control

over this content over to FB.[28] Paradoxically, FB remains unaccountable for the content since it isn't the author. Secondly, FB uses all content and other personal data from online surveillance as the factual payment for a user's FB voice. User data has a market value that most users are not aware of. So, a FB voice is not "free", but the actual value FB turns personal data into is unknown to the user. Thirdly, as argued, all content on FB is moderated, so a user's online voice and social network are not a reflection of unmoderated reality.

If, despite these serious downsides, we would still wish to maintain the goods of innovation and user access to online expression, we need to let go of the idea that social media should implement the HRFE. It is impossible to adjust the HRFE to self-regulated social media.

Let me quickly sketch why this is an impossibility. Since currently the scope of FE online is smaller than its scope in offline liberal democracy, the solution would need to be to reduce the scope of the HRFE to reflect online expression.[29] What would we attain by doing that? We would attain the dubious claim that online expression is congruent with the HRFE. To maintain that claim we would need to adjust the HRFE every time FB adjust its content moderation policies. Therefore, in effect, it would extent FB's self-regulation to include the regulation of the HRFE. This would be either impossible or it would effectively nullify the HRFE.[30] Also, on this scenario, the HRFE would still not be a claimable right on social media.

4.1 Some ideas on regulation of social media

I argued that there is no gain in adjusting our conception of FE to self-regulated social media. Instead, liberal democracies should regulate social media into adhering to FE in the legal form of the HRFE.

---

[28] This includes private messages sent through FB's Messenger service. FB's Statement of Rights and Responsibilities states that, "[f]or content that is covered by intellectual property rights, like photos and videos (IP content), you specifically give us the following permission, subject to your privacy and application settings: you grant us a non-exclusive, transferable, sub-licensable, royalty-free, worldwide license to use any IP content that you post on or in connection with Facebook (IP License). This IP License ends when you delete your IP content or your account unless your content has been shared with others, and they have not deleted it" (Facebook, 2018).

[29] There is some irony in the fact that FB's self-regulation is used to reduce FE instead of reducing its regulation.

[30] Offline expression would lose its validity as a protection of democratic legitimacy, in part because the scope of the HRFE would be based on business concerns rather than liberal democratic concerns. The other consequence it fails to address is that, if online expression becomes determinative for the HRFE both online and offline, its fluctuation in scope could go either way, towards greater restriction and towards greater freedom, based on innovation in online reality and not taking into account offline reality. Since social media like FB would not alter their business structure, users would not be able to claim even this new online HRFE as a right. It would only effectively nullify their offline HRFE and increase their powerlessness in relation to FB online. Business revenue would trump human rights.

Someone may object that the traditional account of FE is outdated and unsuitable for being applied to social media. But, since the traditional account of FE defines the conditions to adhere to or to fulfill, for FE to be able to function as a protector of liberal democracy, and, if liberal democracy is the political structure we want, then the traditional account of FE is normative, also for expression on social media.

Regulation of social media by liberal democratic governments

The most obvious answer for liberal democracies is that its governments should regulate online expression to protect the national conception of the HRFE of their citizens, also on social media. But current legal structures and FB's globalization of communication do not make this solution as practically feasible as it seems obvious. The questions are, how can government regulation protect FE without violating Facebook's FE in the process? And how do we impose the reality of a nationally defined and physical society on a reality that is global and nonphysical? Lastly, how about the downsides of governmental regulation?

First, let's consider how government regulation can protect FE on social media without violating FB's FE in the process. American private corporations like FB Inc. are legally considered to be a person, with a legally protected FE in the form of the First Amendment (CRS, 2019). The employees and customers of private companies have the same legal protection defined in the First Amendment. But the First Amendment protects people against state interference, not against interference from private companies like FB Inc. Therefore, FB Inc. is protected against government interference with its free speech rights, but the company does not need to respect the free speech rights of its users.

Government regulation then, could consist of implementing law necessary to secure the First Amendment rights of people, also against social media companies. But, as the Congressional Research Service [CRS] points out, government regulation of social media is likely to be legally interpreted as infringement on the First Amendment rights of the company (CRS, 2019, p.4).

Still, the CRS presents three possible perspectives from which to consider government regulation on social media. By "[l]ooking to three possible analogues drawn from existing First Amendment law, the report explores whether social media companies could be viewed in the same way as company towns, broadcasters, or newspaper editors" (CRS, 2019, p.4).

The first possible analogy, "*Company towns*", figured in an American Supreme Court ruling that stated that, in special cases, "private actors should be treated as the government and must comply with constitutional standards when interacting with others" (CRS, 2019, p.23). The

name refers to such a special case concerning a company-owned town (a town, functioning like other towns, but owned by a private company) where citizens got First Amendment protection (CRS, 2019, p.23). The argument for treating social media after the precedent of the company-town case is that social media platforms have a predominantly public function and can be likened to a public square (CRS, 2019, pp.24-5).

In the second possible analogy, broadcasters are legally treated as "common carriers" (i.e. historically, "an entity that holds itself out to the public as offering to transport freight or passengers for a fee") and form an industry subject to government requirements and regulations (CRS, 2019, p.27). Social media could be treated as legally analogous to common carriers, and therefore legitimately subject to government regulation (CRS, 2019, p.27).

In the last possible analogue, newspaper editors, when selecting or editing expressions before publishing, legally engage in a speech act that is protected by their First Amendment Rights (CRS, 2019, p.33). Moderating content in the manner social media do may thus be argued to be legally equivalent with editing and therefore protected by free speech laws (CRS, 2019, p.33). But this would give tech companies the status of media company, which means different laws apply.

These are some suggestions the CRS gives concerning legal forms government regulation of social media might take. But how do we impose the reality of a nationally defined and physical society on a reality that is global and nonphysical? The contrast between traditional FE and expression on social media gets strengthened by the fact that they treat expressions so differently. Governmental regulations resulting in an integrated account of FE need to give a clear account of global communication and how its forms and consequences relate to national realities. A challenge is that regulation by national governments ties FB to different nationally determined legal implementations of the HRFE. Seen from the perspective of FB Inc., a more global solution would seem more relevant than many different nationally determined ones.

Another concern are the downsides of government regulation of social media. Historically, governmental regulation has been more prone to censoring of unwanted political voices. Government regulation can result in online expression not being a free haven for (political) expression. An extreme example of this, but of a (not liberal democratic) communist government, is China. In China, tech companies seeking to be an online platform in the country must censor all content which the Chinese government considers "sensitive" (Kharpal, 2019). Though tech companies still look for business opportunities in China, none have speech related online platforms in the country. Instead, China owns its own international social media platforms such as "WeChat" and "TikTok".

Other regulators?

These challenges might bring up the question of who the regulator of social media should be. Are there other entities than national governments that could play the part?

Could the tech companies that facilitate the social media platforms ratify relevant parts of the Universal Declaration of Human Rights [UDHR] directly? This would solve the challenge of determining what online expressions belongs to which nationally determined interpretation of the HRFE. But the solution seems to run into problems of enforceability. Without going into details about what constitutes a government, it can easily be argued that, since FB Inc. is not a government, it therefore cannot realistically enforce the HRFE. The accountability of enforcing the UDHR online can likely not be moved back a factor to the UN either.

Alternatively, maybe there should be a global democratic organ for the regulation of online expression. Though likely this organ would still need to be grounded in the national governments of all countries involved. Investigating such scenarios in depth are beyond the scope of this thesis but even if the UN or a new global organ could successfully regulate online expression, there would arise a new form of the challenge of solving discrepancies between national and international conceptions of the HRFE dividing expressions offline and online.

Tech companies might propose a hybrid solution in the form of voluntary abidance by the HRFE of tech companies with the help of nationally determined enforceability. This would mean tech companies would need to face the challenge of bridging the gap between online expression and each nationally determined HRFE. In turn this would provide them with each government's support with determining and enforcing the scope of online expression. National governments would need to address all aspects of where legal frameworks are lagging behind online reality. But tech companies would need to alter their business model on the hybrid scenario. The HRFE would have to trump business concerns where the two conflict, and it is unlikely that social media companies would do this voluntarily because of the losses this entails.

Another hypothetical scenario to consider would be if FB users collectively (as a kind of online *demos*) demanded a traditional conception of FE as a claimable right from FB Inc. directly. Then FB would maintain its independence from external governmental regulation.

But FB users would want a rights claim on the company that would require fundamental changes to the businesses' policies and structure. According to my earlier argument, FB would need to stop interfering with users' harmless expressions and face the challenge of prioritizing FE and protection from harm over business considerations. The only upside for FB would be its continued self-regulating in relation to liberal democracies. This scenario leaves unaddressed

the question of what relation would exist and be optimal between liberal democracies and expression on social media.

Deciding on a practical solution for how to solve the discrepancies between traditional FE and online expression is beyond the scope of this thesis. But this thesis aims to show the extent and nature of the gap that the rapid evolution of social media has left for offline liberal democratic society to catch up with. If we want to maintain liberal democracy as the basic structure of society, FE (and its legal form as the HRFE) plays an essential role in protecting that political structure. Therefore, liberal democracies need to find a way to include or account for FE on social media.

5 Conclusion

Should we rethink the Human Right to Freedom of Expression now that social media have become such a dominant part of individual expression and public discourse? Article 19 of the UDHR states that

> Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers (UDHR, 1948).

The HRFE means to secure a universal right, for every human being, to have the freedom to think, share, hear, question, and read all sources of knowledge and information, regardless of person or place. The traditional account of FE acknowledges the legitimacy of restriction of this right to protect others from harm.

In this thesis I have, with the help of a philosophical normative foundation and real-life cases regarding social media, argued the importance of the HRFE for liberal democracy. From my discussion it becomes clear that it is not possible for a user on social media like FB to enjoy the HRFE as a claim right. Legally, this conclusion is unproblematic, since social media have no legal duty to secure its users' HRFE. But I argue that this conclusion is problematic from the philosophical view which stresses the importance of FE for liberal democracy.

I started this thesis with a discussion of the legal content and form of the HRFE. Essential to the legal content is the universality of the human right, both concerning each human being as well as regarding all media and any geographical location. The intention of the legal conception of the HRFE thus includes social media.

In the second chapter I discussed the philosophical foundation for freedom of expression. Mill's account is central to our understanding of what role FE plays in liberal democracy, how it organizes and protects civil liberty in relation to its governing power, and how it is meant to protect society against threats from tyrannical majority opinion or an assumption of infallibility from the government.

In addition to Mill, I discussed the views of Dworkin and Waldron on what legitimizes restrictions of FE. Dworkin's argument is that the free speech principle should be kept intact for it to legitimize democratic decisions. The protection of citizens from harm should be delegated to other aspects of the legal system. This prioritization of the liberal aspect of liberty seems to insufficiently account for the harm caused by harmful misinformation and maliciously

manipulative content on social media. Waldron on the other hand, prioritizes equality over liberty by arguing for regulation of FE to protect minorities from hate speech. But such protection may start a slippery slope into a government's assumption of infallibility. This objection is illustrated by FB's arbitrary content moderation policies.

In chapter three I analyzed aspects of FB in relation to the traditional account of FE, FB being a representative case for social media. I distilled the traditional account of FE into five conditions, namely: the basic condition that the expression is not harmful; one person, one voice; the necessity of openness to dialogue of honest speakers; real and morally accountable people; and the offline conception of a closed, physical society.

The discussion showed that, though FB has an intention to make possible online expression for as many people as possible, the platform interferes with the enjoyment of its users' HRFE in all other aspects mentioned in article 19 of the UDHR. In addition, prioritizing technological innovation and business revenue, FB addresses problems it causes after the harm has been done.[31] The speed and reality of online innovation has created a schism between offline and online expression that makes online expression a constant violation of FE and leaves offline realities with far-reaching violations of democratic processes to clean up.

Since FE plays such an essential role in expressing and protecting liberal democracy in society, I believe that its legal form of the HRFE should be ensured in, at least, all domains in which public discourse takes place. Social media host a large and important part of national and international public discourse. Therefore, to respect and ensure the proper function of FE in liberal democracy, social media should abide by the HRFE.

My discussion of FB as a representative case shows that, with social media self-regulating independently from liberal democratic society and the consequent absence of FE online, the two Millian threats do take place on social media. Therefore, Mill's point of the dangers of a society without FE is illustrated by the neglect of FE on social media like FB.

In chapter four I discuss how we could harmonize the discrepancies between the traditional account of FE and social media. This discussion lands us in (legal and political) practical concerns that are beyond the scope of this paper to properly address or solve. But the discussion shows the current gap between the traditional FE in liberal democracies and online expression on social media. I conclude chapter four with short sketches of what forms social media regulation might take.

[31] For many years, FB's motto was "move fast and break things".

The philosophical contribution this thesis purports to make to the account of the HRFE in the age of social media, is the awareness that currently expression on social media is divorced from our traditional account of FE and that this is harming liberal democracies.

My claim is that, for traditional FE to achieve what it is intended to achieve, expressions on social media should be regulated in accordance with the HRFE. I believe that the most obvious solution would be for nations to be able to regulate FB's content policies as they apply within their borders. I believe this should be so because social media host an enormous part of public discourse essential for liberal democracy, because a substantial part of the population of every liberal democracy uses social media as a platform for expression, and because expression on social media impacts liberal democratic society and its political processes and public opinion in essential ways.

6 Literature

Alexander, Larry (2003). Freedom of expression as a human right. In *Protecting Human Rights* (pp. 39-73). Oxford: Oxford University Press.

Allan, Richard (2018/August 9th). Hard Questions: Where Do We Draw the Line on Free Expression? In *Facebook*. Retrieved from https://newsroom.fb.com/news/2018/08/hard-questions-free-expression/.

Amer, Karim & Noujaim, Jehane (2019). *The Great Hack* [documentary]. USA: Netflix.

BBC (2015, March 12th). Ferguson Police shot during protest. In *BBC News*. Retrieved from https://www.bbc.com/news/world-us-canada-31846425

BBC (2020a, February 10th). Mila: Teen who criticised Islam on Instagram moves school. In *BBC News*. Retrieved from https://www.bbc.com/news/world-europe-51446519

BBC (2020b, May 4th). Covid-19: Investigating the spread of fake coronavirus news. In *BBC News*. Retrieved from https://www.bbc.com/news/av/technology-52477361/covid-19-investigating-the-spread-of-fake-coronavirus-news?SThisFB&fbclid=IwAR3zB7by6Co84T_02gbfd9MqzkuzdPIpjaVfQgkJQuRKajf04okjuIViWjU

Begley, Ian (2017, May 23rd). Storm over Facebook failure to block live online self-harming. In *Independent.ie*. Retrieved from https://www.independent.ie/business/technology/news/storm-over-facebook-failure-to-block-live-online-self-harming-35744400.html

Boot, Eric R. (2017). Two Problems with Limitless Freedom. In *Human Duties and the Limits of Human Rights Discourse* (pp. 147-150). Cham (SW): Springer International Publishing.

Bowers, John; Zittrain, Jonathan (2020). Answering Impossible Questions: Content Governance in an Age of Disinformation. *The Harvard Kennedy School (HKS) Misinformation Review*, Volume 1, Issue 1.

Brink, David (2001). Millian principles, freedom of expression, and hate speech. *Legal Theory, 7* (2001), pp. 119–157.

Brink, David (2018). Mill's Moral and Political Philosophy. In *The Stanford Encyclopedia of Philosophy (Winter 2018 Edition)*. Edward N. Zalta (ed.). Retrieved from https://plato.stanford.edu/archives/win2018/entries/mill-moral-political/

Cadwalladr, Carole (2020, January 4th). Fresh Cambridge Analytica leak 'shows global

manipulation is out of control. In *The Guardian*. Retrieved from https://www.theguardian.com/uk-news/2020/jan/04/cambridge-analytica-data-leak-global-election-manipulation

CBSN (2018, April 11th). Facebook CEO Mark Zuckerberg faces second day of testimony. In *CBS News*. Retrieved from https://www.cbsnews.com/video/facebook-ceo-mark-zuckerberg-faces-second-day-of-testimony/

Communications Decency Act 1996 (USA). Retrieved from https://www.minclaw.com/legal-resource-center/what-is-section-230-of-the-communication-decency-act-cda/

Children's Online Privacy Protection Act 1998 (USA). Retrieved from https://www.ftc.gov/enforcement/rules/rulemaking-regulatory-reform-proceedings/childrens-online-privacy-protection-rule

Congressional Research Service. (2019, March 27th). *Free Speech and the Regulation of Social Media Content* (R45650). Retrieved from https://fas.org/sgp/crs/misc/R45650.pdf

Convention on the Rights of Persons with Disabilities. (2007). Convention on the Rights of Persons with Disabilities. Retrieved from https://www.ohchr.org/EN/HRBodies/CRPD/Pages/ConventionRightsPersonsWithDisabilities.aspx#21

Cyphers, Bennett (2019). A Guided Tour of the Data Facebook Uses to Target Ads. On *Electronic Frontier Foundation*. Retrieved from https://www.eff.org/deeplinks/2019/01/guided-tour-data-facebook-uses-target-ads

Donatelli, Piergiorgio (2006). Mill's Perfectionism. In *Prolegomena* 5 (2) 2006: 149–164.

Donath, Judith (1995). Identity and deception in the virtual community [draft]. Retrieved from https://www.researchgate.net/publication/2512169_Identity_and_Deception_in_the_Virtual_Community

Donath, Judith (2019). Artificial Entities [video]. In: *Toward a handbook of Ethics of AI*, an interdisciplinary workshop hosted by the Centre for Ethics, University of Toronto, March 1-2, 2019. Retrieved from https://www.youtube.com/watch?v=QGHONm1xLz0&fbclid=IwAR2o8lVDDDeKu1XEwyJ6PlS1eIgnhpGN1RFgNpCpBWSZaehH7WIaI1H9jUg

Dworkin, Ronald (1997). *Taking Rights Seriously*. London: Duckworth.

Dworkin, Ronald (2006). A New Map of Censorship. In *Index on Censorship*, pp. 130-133.

https://doi.org/10.1080/03064220500532412

Dworkin, Ronald (2009). Foreword. in *Extreme Speech and Democracy* (pp. iii-ix). Editors: Ivan Hare, James Weinstein. Oxford: Oxford University Press.

Dworkin, Ronald (2012). Session 4: Multiculturalism and Human Rights [video]. At *Challenges to Multiculturalism. A Conference on Migration, Citizenship, and Free Speech*, 25-26 June 2012 at The House of Literature in Oslo. Retrieved from https://www.youtube.com/watch?v=6wJQ658e-4U

Economist, The (2017, May 6th). Regulating the internet giants. The world's most valuable resource is no longer oil, but data. In *The Economist.* Retrieved from https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

European Convention on Human Rights. (1950). The Convention for the Protection of Human Rights and Fundamental Freedoms. Retrieved from https://www.echr.coe.int/Documents/Convention_ENG.pdf

Epstein, R. & Robertson, R. E. (2015). The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. In *PNAS August 18, 2015* 112 (33) E4512-E4521. Retrieved from https://doi.org/10.1073/pnas.1419828112

Facebook (2018, January 31st). Statement of Rights and Responsibilities. In *Facebook.* Retrieved from https://www.facebook.com/legal/terms/previous

Facebook (2019a). Company info. In *Facebook*. Retrieved on June 30th, 2019, from https://newsroom.fb.com/company-info/

Facebook (2019b). Community Standards. In *Facebook*. Retrieved on July 1st, 2019, from https://www.facebook.com/communitystandards/introduction

Facebook (2019c). Oversight Board Charter. In *Facebook*. Retrieved from https://fbnewsroomus.files.wordpress.com/2019/09/oversight_board_charter.pdf

Facebook (2020). Terms of service. In *Facebook.* Retrieved on March 17th, 2020, from https://www.facebook.com/legal/terms

Frankfurt, Harry G. (2005). *On Bullshit*. New Jersey: Princeton University Press.

Frontline (2018). *The Facebook Dilemma* [video]. Retrieved from https://www.pbs.org/wgbh/frontline/film/facebook-dilemma/

General Data Protection Regulation. (2018). General Data Protection Regulation. Retrieved from https://gdpr.eu/

Gilabert, Pablo (2009). The feasibility of basic socioeconomic human rights: a conceptual exploration. In *Philosophical Quarterly, October 2009*, Vol.59(237), pp.659-681.

Go, Johann (2018). Mill and the Limits of Freedom of Expression: Truth, Lies, and Harm. In *International Journal of Applied Philosophy, 2018*, Vol.32(1), pp.1-18.

Gorman, Ginger (2019). *Troll hunting*. Melbourne: Hardie Grant Publishing.

Grinberg, Emanuella (2014). Facebook 'real name' policy stirs questions around identity. In *CNN.* Retrieved from https://edition.cnn.com/2014/09/16/living/facebook-name-policy

Hannikainen, Lauri and Myntti, Kristian (1993). Article 19. In Asbjørn Eide (Ed.) *The Universal Declaration of Human Rights: a commentary* (2nd ed.) (pp. 275-286). Oslo: Universitetsforlaget.

Hobbes (1996). *Leviathan*. Richard Tuck (Ed.). Cambridge: Cambridge University Press.

Hohfeld, Wesley Newcomb (1913). Some Fundamental Legal Conceptions as Applied in Judicial Reasoning. *The Yale Law Journal, 1 November 1913*, Vol.23(1), pp.16-59.

House of Commons, Digital, Culture, Media and Sport Committee (2019). *Disinformation and 'fake news': Final Report* (Eighth Report of Session 2017–19). Retrieved from https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/1791.pdf

Jackson, Benjamin F. (2014). Censorship and freedom of expression in the age of Facebook. *New Mexico Law Review*, 44(1), pp. 121-168.

Kelly, Heather (2019, May 15th). Facebook changes livestream rules after New Zealand shooting. In *CNN Business*. Retrieved from https://edition.cnn.com/2019/05/14/tech/facebook-livestream-changes/index.html

Kharpal, Arjun (2019, July 17th). Google has been accused of working with China. Here's what they've been doing there. In *CNBC Tech*. Retrieved from https://www.cnbc.com/2019/07/17/google-china-what-businesses-the-search-giant-has-in-the-country.html

Kierulf, A., Gisle, J. & Elden, J. C. (2018). Ytringsfrihet. In *Store norske leksikon på snl.no*. Retrieved on May 19th, 2020, from https://snl.no/ytringsfrihet

Kohlberg, L., Levine, C. & Hewer, A. (1983). *Moral stages: a current formulation and a response to critics*. Basel, NY: Karger.

Kozlowska, Hanna (2018). Facebook's fight against bad content is a mess. In *Quartz*. Retrieved from https://qz.com/1329967/facebooks-fight-against-bad-content-is-as-chaotic-as-ever/

Mill, John Stuart (1989). *On Liberty*. Stefan Collini (Ed.). Cambridge: Cambridge University Press.

Miller, David (2003). *Political Philosophy: A very short introduction*. Oxford: Oxford University Press.

Mosseri, Adam (2016, June 29th). Building a Better News Feed for You. In *Facebook*. Retrieved from https://newsroom.fb.com/news/2016/06/building-a-better-news-feed-for-you/?ref=fbb_blog

Mosseri, Adam (2018, January 11th). News Feed FYI: Bringing People Closer Together. In *Facebook*. Retrieved from https://www.facebook.com/business/news/news-feed-fyi-bringing-people-closer-together

Mozur, Paul (2018, Oct. 15th). A Genocide Incited on Facebook, With Posts from Myanmar's Military. In *New York Times*. Retrieved from https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html

Nelson, Joshua (2020, May 27th). Judge Napolitano: 'Twitter has freedom of speech just like you and I and the president himself'. In *Fox News*. Retrieved from https://www.foxnews.com/media/judge-andrew-napolitano-twitter-fact-check-trump

Nussbaum, Martha C. (2004). Mill between Aristotle & Bentham. *Daedalus*, 133(2): pp. 60–8. Retrieved from https://doi.org/10.1162/001152604323049406

Pettit, Philip (2011). The Instability of Freedom as Noninterference: The Case of Isaiah Berlin. *Ethics*, Vol. 121, No. 4 (July 2011), pp. 693-716.

Quinn, Ben (2020, May 1st). Facebook removes page belonging to conspiracy theorist David Icke. In *The Guardian*. Retrieved from https://www.theguardian.com/media/2020/may/01/coronavirus-facebook-removes-page-conspiracy-theorist-david-icke?CMP=Share_AndroidApp_Correo

Raz, Joseph (1986). *The morality of Freedom.* Oxford: Clarendon.

Roose, Kevin (2019, August 4th). 'Shut the Site Down,' Says the Creator of 8chan, a Megaphone for Gunmen. In *The New York Times*. Retrieved from https://www.nytimes.com/2019/08/04/technology/8chan-shooting-manifesto.html

Sage, Adam (2020, January 24th). French teenager in hiding after insulting Islam online. In *The Times*. Retrieved from https://www.thetimes.co.uk/article/french-teenager-in-hiding-after-insulting-islam-online-0vl5hrs0m

Ten, C.L. (1998). Democracy, socialism, and the working class. In: *The Cambridge Companion to Mill* (pp. 372-395). Edited by John Skorupski. Cambridge: Cambridge University Press.

Universal Declaration of Human Rights. (1948). The Universal Declaration of Human Rights. Retrieved from https://www.un.org/en/universal-declaration-human-rights/

United States Constitution, Amendment I.

Waldron, Jeremy (2012). *The harm in hate speech*. Cambridge, MA: Harvard University

Press.

Washington Post, the (2019, July 24th). FTC chairman on Facebook fine: 'The enormity of this penalty resets the baseline for privacy cases' [video]. In *the Washington Post.* Retrieved from https://www.washingtonpost.com/video/business/technology/ftc-chairman-on-facebook-fine-the-enormity-of-this-penalty-resets-the-baseline-for-privacy-cases/2019/07/24/c6bbf8db-76d0-44bd-8ab1-520f3b2678d3_video.html?

Werhan, Keith (2008). The Classical Athenian Ancestry of American Freedom of Speech. In *The Supreme Court Review*, Vol. 2008, No. 1 (2008), pp. 293-347. Chicago: The University of Chicago Press

Zuckerberg, Mark (2018, January 12th). (No title) [Facebook post]. In *Facebook*. Retrieved from https://www.facebook.com/zuck/posts/10104413015393571

Zuckerberg, Mark (2019a, June 27th). (No title) [video]. In *Facebook*. Retrieved from https://www.facebook.com/zuck/videos/10107820049450011/

Zuckerberg, Mark (2019b, October 17th). Standing for Voice and Free Expression. In *Facebook*. Retrieved from https://about.fb.com/news/2019/10/mark-zuckerberg-stands-for-voice-and-free-expression/