A Crisis of Liberalism: Misinformation Regulation on Social Media and the Future of Digital Freedom of Expression

by Margaret Holloway

MA Human Rights Thesis
Supervisor: András Sajó
Legal Studies Department at Central European University

Vienna, Austria
June 2021

Table of contents

Abstract:

The task of regulating online speech is a contemporary and future governance challenge for both public and private actors. The digital public sphere is a relatively recent phenomenon and is constantly and quickly evolving; as a result, questions of regulation are difficult to thoughtfully answer, and the actions of social media companies in regulating content easily seem reactionary rather than grounded in guiding political theories of expression. This thesis aims to answer the following question: *how does a liberal theory of freedom of expression apply to the world of digital speech?* It begins with an overview of John Stuart Mill's theory of freedom of expression and of the liberal political goals of maintaining a plural, open society and preserving individual autonomy and independence from authority. The first chapter also includes a doctrinal analysis of the liberal *harm principle* and its theoretical issues. The second chapter presents the practical issue of regulating misinformation on social media. It explores the political dimension of misinformation and evaluates Facebook's content regulation policy and practices against traditional liberal standards, particularly the harm principle. Ultimately, the author argues that the various approaches already taken to tackle misinformation on social media do not fall in line with traditional liberal theory, and that the future of a liberal culture online will require increased focus on protecting individual data privacy, deprioritizing commercial interests, and building truly pluralistic public platforms that encourage critical thinking.

Introduction: Liberalism and expression in the digital era

We live in an era of rapid and mass communication, where the "digital public" increasingly merges with the notion of "public" altogether. Liberal democratic states are fundamentally structured by a distinction between the public, or the "people", and the government; the ongoing but sustaining tension of a liberal democracy is the issue of finding a proper balance of power and control between the government and the public. This pursuit is complicated by conceptual difficulties in distinguishing between the "people" and the government, when the form of government (democracy, 'rule of the people') would suggest that the people rule themselves. Political theories of liberalism acknowledge this conceptual difficulty and suggest that the principle of liberty should protect individuals and their acts from government interference, as the existence of a higher authority over the public is non-negotiable, even in a democracy. Accordingly, as the digital era ushers in changes to the shape and functioning of the public sphere, which is the stage of our acts as individuals, we are called to reflect upon the function of the government in regulating it (or leaving it be).

Although traditional liberal political theory is broad and intends to provide a framework for thinking about the entire project of governing society, it is largely based on the more specific concepts of individuality and private life and belief. Francis Fukuyama, notable for declaring the "End of history" along with the triumph of liberal democratic ideology after the Cold War, simply describes liberalism as "a system for peacefully managing diversity in pluralistic societies."[1] He poetically notes that liberalism has "sought to lower the temperature of politics by taking questions of final ends off the table and moving them into the sphere of private life", and that it "thus protects diversity by deliberately not specifying higher goals of human life."[2] In other words, liberalism enables the sustainability of social pluralism by establishing a carefully calculated distance between the individual and the government. In this imaginary, the sphere of politics is easier to navigate because it purposefully does not take up "unsolvable" questions: those which engage individuals and social groups so deeply that genuine compromise between them is unlikely or impossible.

In the digital era, new norms of communication and information-sharing mean we are all constantly flooded with the opinions of others. We see, hear, and read more of those ideas which we agree with and incorporate into our own worldview, but also more of those which

[1] Francis Fukuyama. "Liberalism and its discontents: The challenges from the left and the right," *American Purpose,* 5 Oct. 2020. https://www.americanpurpose.com/articles/liberalism-and-its-discontent/

[2] Ibid.

we oppose. In this way, the digital public sphere may heighten our own sense of conviction, as it allows us to develop our opinions at the same time that reveals to us the existence of those on “the other side” of the ideological divide encouraged by the polarization of the media and the political discourse it shapes. I do not mean to say that the digital public sphere encourages a bi-polar mode of political thought, but rather one that breaks off into many fractions and at the same time “raises the temperature” of politics. In the end, this creates a political world where we are enabled and encouraged to think and feel deeply about the possibility of truths and final ends in society, but are still expected to keep our convictions largely private.

A focus on speech and expression is particularly useful in assessing the way in which liberal doctrine applies to governance in contemporary societies. Digital technology has shaped our common world in a way that involves all parts of society. The accessibility of information and far-reaching speech platforms means that the individual, in both private and public (or political) life, has a fundamentally new way of relating to others and to authority. Because liberal doctrine provides a framework for regulating exactly these relationships between the individual and society, it follows that the applicability of the theory will shift along with changes to the structure of society.

At the core of liberal theory is the notion that the individual may freely develop themselves without interference from the state, so long as they are not harming others.[3] The act of expression is intrinsically such an act of self-development, as the communication and formation of opinion is the foundation of decision-making. Judith Butler reflects on how we as subjects are constituted by language; without the right to free expression, the process of self-development could not even begin.[4] Through the act of expression, too, we see ourselves as both an individual and part of society. Expressing our ideas is for us a means of placing ourselves as one particular individual among the many others around us. The digital era has expanded the format for expression and has also expanded the range of potential effects of our expression (on ourselves and others). In this way, the concept of individuality has changed, and the boundary between individual and society, which determines appropriate levels of government interference, has perhaps also shifted.

[3] This is John Stuart Mill’s *harm principle*; it will be a focus later.

[4] Judith Butler, *Excitable Speech: A Politics of the Performative*, (New York & London: Routledge, 1997).

Chapter One: The freedom and limits of expression in traditional liberal theory

I. Liberty in J.S. Mill's "On Liberty"

a. Liberty in the democratic state

J.S. Mill's *On Liberty* is a foundational source of liberal thought; despite its wide influence, the text's framework around the concept of liberty is nuanced and is still the subject of significant debate in legal and political theory. Nonetheless, Mill's text is particularly applicable to the liberal democratic model of government because it begins with an explanation of the role of liberty within the particular context of the democratic state. The connection between liberalism and democracy may seem self-apparent or intuitive at the current point in history, where the political ideals of liberal democracy seem so well-defined. However, it is important here to take a step back and consider that the individualistic impulse of liberalism and the collectivity that is so important to democracy means that the two are perhaps not naturally connected. The disjuncture between liberalism and democracy, in fact, is Mill's justification for a focus on preserving and protecting individual liberty.

Mill answers an important question which reaches towards the conceptual tension between liberty and democratic government: why is it so important to protect individual liberty in a state where the form of government (democracy) would suggest that individuals are already involved in self-government, which is in some sense a form of liberty itself? Can we not assume that a government principled on the concept of self-rule will automatically allow the individual space to exist outside of the influence of the state? Mill would suggest not, and his reasoning here also suggests that the principle of liberty in some sense fulfills the rhetorical promises of democracy: that is, the prioritization of the individual's will and autonomy. So, without a concurrent focus on the protection of individual liberty, democracy easily forgets about its own basic emphasis on the individual. The individual's political will is essential to democracy; the rule of the "people", or the majority, can only come into form as a collective of individual wills. In other words, the value of democracy as a means to legitimize and follow the will of the people depends on the ability of individuals to freely form and express individual opinions, and then to freely communicate and organize these ideas to form a collective opinion. However, there is a significant tendency for both government power and social majorities to become coercive to the minority opinion, which disturbs the balance between individual and the society which allows for the democratic state to persist as a truly representative system.

Put differently, Mill reinforces that the concept of individual liberty is essential to democracy, but challenges the assumption that democracy as a political system automatically protects it. Mill explains that:

> It was now perceived that such phrases as ‘self-government’, and ‘the power of the people over themselves’, do not express the true state of the case. The ‘people’ who exercise the power are not always the same people with those over whom it is exercised; and the ‘self-government’ spoken of is not the government of each by himself, but of each the rest.[5]

Democracy, then, is a system which should begin with a framework for protecting liberty. The democratic state’s willingness to rule according to the will of the people should not preclude its protection of those whose ways of believing and acting in the world diverge from the rest. Such oppression of the minority will quickly defeat the democratic project altogether; in such a state, neither society nor politics could progress except within the terms of the already dominant.

b. The “tyranny” of the majority & the marketplace of ideas

The well-known concept of the “tyranny of the majority” is central to modern concepts of liberalism as informed by Mill. It is also central to liberal justifications for freedom of expression; if dissenting voices are silenced or intimidated before they have the chance to speak, the democratic system is not necessarily destabilized but it begins to lose legitimacy. Here, legitimacy is tied to the independence of opinion; if expression is unequal in the public sphere, then individuals do not have, in Mill’s mind, the true freedom of opinion formation. In calling such a democracy illegitimate, I mean that it no longer exists as a truly representative government; while it may still represent the opinion of the majority or the “people”, it has ceased to be a society which preserves the public sphere as a platform for the free expression of ideas, and therefore the free development of opinions and beliefs. The existence of a social or political majority correctly implies that there will always be parts of society which hold fast to opinions different from those of the rest.

> Society can and does execute its own mandates: and if it issues wrong mandates instead of right, or any mandates at all in things with which it ought not to meddle, it practices a social tyranny more formidable than many kinds of political oppression, since, though not usually upheld by such extreme penalties, it leaves fewer means of escape, penetrating much more deeply into the details of life, and enslaving the soul itself.[6]

[5] John Stuart Mill, “On Liberty,” in *On Liberty, Utilitarianism and Other Essays*, ed. Mark Philip and Frederick Rosen (New York: Oxford University Press, 2015), 7.

[6] Mill, “On Liberty,” 8.

The language of “tyranny” is unique to Mill’s understanding of liberalism and democracy; in democratic theory that falls further from liberal doctrine, the rule of the majority opinion is not necessarily cast in the same negative light as it is here. This difference in tone helps us understand the deep message of doctrinal liberalism, which is that individual liberty is a primary value, and that unjust influence from the society is not potentially, but inherently, threatening. This form of “tyrannical” control, which may be exercised by a government or by “private” society, is dangerous precisely because its majoritarianism lends it, by some standards, an image of legitimacy. However, just because such control is bottom-up in form, rather than top-down (as we imagine most tyrannical governments), does not mean that there is a free environment. Mill’s language in the passage above is admittedly vague and perhaps problematic; he does not specify what may constitute “things with which it [society] ought not to meddle”, and the notion that mandates may be “wrong” or “right” may appear to mirror precisely the forms of control which he warns against. In order to avoid the arbitrary application or acceptance of this theory then, I will proceed by considering later in my writing the particular example of expression-related mandates, against the wider backdrop of expression in the democratic society.

Along with the “tyranny of the majority”, Mill’s notion of “the marketplace of ideas" sits at the base of liberal thought, and is particularly central in the context of the liberal approach to expression. The language of the “marketplace of ideas” is actually not included in Mill’s text; rather, it is a metaphor which serves as the conclusion of Mill’s reasoning in Part II. of *On Liberty*, “Of the liberty of thought and discussion”. The free-market principle of competition is applied to the world of ideas; in extremely simple terms, in the “marketplace of ideas” the truthfulness of ideas can never be assumed as a given, and only an environment of free expression can truth develop, emerge, and ultimately “win out” over false ideas. Of course, the issue is more complex than stated this way. Importantly, the metaphor’s emphasis on the triumph of truth may be misleading. Mill never reaches the conclusion that free expression will lead to the discovery of *the* coherent truth, and therefore also does not imply that the discovery of objective truth should be a goal of society, or at least a goal that free expression can ever reasonably achieve. What Mill does suggest is that all opinions may contain partial truths, and that partially or wholly truthful opinions are only meaningful and useful in society when they are held with a critical attitude. The expression policy resulting from this reasoning is one where “truthfulness” is not a valid criterion along which to permit or limit speech. In other words, false opinions cannot be censored because they are false.

The basic condition of plurality is essential to both concepts mentioned here (tyranny of the majority; marketplace of ideas). The democratic state is not legitimate without the possibility to oppose the majority; an opinion or idea is not legitimate, as in assuredly true, if it is not contested and defended. Recalling that liberalism is in some sense a means of making pluralistic societies politically tenable, the openness of the liberal approach to expression may come as a surprise (given the politically difficult nature of some expression). Put differently, assuming that a high level of social pluralism will lead inevitably to some political conflict, it may seem like a fool's errand to create a system of governance which seeks to avoid high-level conflict at the same time that it fosters the pluralism which sits at the base of conflict. However, from the liberal perspective, free expression *enables* the condition of plurality to persist; it concerns the ability of the individual to maintain his autonomy at the same time that he exists as one of many in the society. The potential limits of free expression, too, are determined by a consideration of the line between the individual in autonomy and in society. Within the liberal doctrine, this line is drawn at the point where the acts of the individual threaten harm to others. The following section will consider this "harm principle" as the boundary of free expression in Mill's work.

c. The harm principle

The harm principle, as derived from *On Liberty*, provides a framework for determining the limits of individual liberty. According to this principle, the individual generally has the freedom to act however he pleases; this freedom, however, ends at the point where an individual's actions cause direct harm to others. This principle is in many ways the thesis of Mill's text. Piers Turner calls the harm principle the "jurisdictional trigger", which asks "the question of what triggers society's jurisdiction."[7] Mill responds to the question of "where does the authority of society begin?" with the following: "As soon as any part of a person's conduct affects prejudicially the interests of others, society has jurisdiction over it, and the question whether the general welfare will or will not be promoted by interfering with it, becomes open to discussion."[8] We can say that, as a member of a larger society, the individual who acts in ways that affect others has somehow forfeited his autonomy; at a certain point, the individualistic concerns of liberalism give way to a more communitarian evaluation of behavior and action. Several questions regarding the application of the principle are immediately apparent: what is harm? how are we to determine the line between indirect and direct harm?

[7] Piers Norris Turner, "'Harm' and Mill's Harm Principle," *Ethics* 124 (January 2014): 306-7.

[8] Mill, "On Liberty," 73.

does intent matter? and so on. The liberal approach may seem to do two conflicting things at once: 1) create a manageable and cohesive society, and 2) allow the individual a broad range of freedoms, both personal and in association with others. Here we come back to the problem of the line between the individual and his society, and to the condition of plurality.

Mill does acknowledge the fluidity of the distinction between the *private* and *public* individual, and therefore the difficulty in determining a threshold of an act's direct influence on others which then warrants interference. The public sphere is defined by social plurality, by the contact, exchange, and interplay between private individuals. Mill admits that "No person is an entirely isolated being; it is impossible for a person to do anything seriously or permanently harmful to himself, without mischief reaching at least to his near connexions, and often far beyond them."[9] So, at what point do an individual's actions affect others to an extent where they may be rightly limited by the society? Does the harm principle then imply that a very wide range of human behaviors may be limited simply because they do, or may, have an effect on others? Mill partially settles the matter:

> But with regard to the merely contingent, or, as it may be called, constructive injury which a person causes to society, by conduct which neither violates any specific duty to the public, nor occasions perceptible hurt to any assignable individual except himself; the inconvenience is one which society can afford to bear, for the sake of the greater good of human freedom.[10]

At best, this reasoning simply reinforces that only *direct* harms can be prevented—that the distant (or contingent) effects of an act cannot justify interference. It does not, however, address the question of what constitutes a harmful effect, or at what level of seriousness the harm becomes intolerable. We are still left to our own devices to determine what inconveniences are so great as to outweigh "the sake of the greater good of human freedom".

This difficulty of the harm principle also concerns morality: although private morality is undoubtedly one of the factors of social pluralism and of the liberal society as free and diverse, the public sphere is intended to be a space somewhat devoid of moral standards, in the sense that public morals not have effect on the individual unless their actions cause harm to others. The opening pages of *On Liberty* include the claim that:

> The object of this Essay is to assert one very simple principle, as entitled to govern absolutely the dealings of society with the individual in the way of compulsion and control, whether the means used be physical force in the form of legal penalties, or the moral coercion of public opinion. That principle is… that the only purpose for which

[9] Mill, "On Liberty," 78.
[10] Mill, "On Liberty," 80.

> power can be rightfully exercised over any members of a civilized community, against his will, is to prevent harm to others.[11]

I would like to point the focus here to the inclusion of "the moral coercion of public opinion" as a form of social control over the individual. This presents a unique challenge to the interpretation and application of the harm principle, which is the difficulty of determining a threshold of harm without involving a moral bias. Given that there are many different forms of harm, which may range from mild to catastrophic, and that the determination of the existence or severity of harm may vary depending on who (or what, in the case of institutional actors) is asked, moral judgments inevitably come into play. If, as an alternative to legal sanction, the "moral coercion of public opinion" is standing eagerly on the other side of the threshold of harm, then it can be said that the determination of harm cannot be made without moral considerations.

Let us return to the notion that liberalism takes the question of "final ends" off of the table of politics and governance, and then consider that the social morality tied up in public opinion is exactly such a "final end" which supposedly cannot be forced upon the individual. Mill argues so strongly "against the interference of the public with purely personal conduct" that it may seem he is also advocating for the expansion of the concept of the individual and private life. He fears ultimately this moral bias of the public against the individual, and the structure of his argument is parallel to that concerning the tyranny of the majority:

> It is easy for any one to imagine an ideal public, which leaves the freedom and choice of individuals in all uncertain matters undisturbed, and only requires them to abstain from modes of conduct which universal experience has condemned. But where has there been seen such a public which set any such limit to its censorship? or when does the public trouble itself about universal experience?[12]

It may follow from an expanded concept of individual and private life that increasingly more harms to society become excusable in the name of protecting greater freedom. What Mill crucially omits from his commentary here is the potential of moral bias in evaluating harm; that is, the subjective nature of what may or may not be seen as harmful. This is, as we will see, a particularly difficult point in considering limits on freedom of expression.

II. The harm in speech

The terms "freedom of speech" (as in the US case) or "freedom of expression" (in Europe and more broadly) themselves are somewhat confusing, as speech that is protected as "free expression" is only free because it falls within certain parameters of acceptable speech.

[11] Mill, "On Liberty," 13.
[12] Mill, "On Liberty," 82.

Of course, an expression policy may actually be quite restrictive; it may limit speech arbitrarily, for purposes of political control, or along the lines of moral and religious sensibility. Within even such a restrictive speech policy, however, unprotected or *unfree* speech exists as an exception to the rule of free expression. So, the language of *free* expression would imply that individual freedom is not deeply or meaningfully affected.

All of this is to say that freedom of expression is not absolute. That is, by itself, largely not a difficult or contentious idea; it makes sense that not absolutely anything can be said at any place at any time (recall the clichéd example of the individual shouting fire in a crowded theater). Still, in practice, the question of what constitutes protected/unprotected speech is complex. The concept of freedom is by its definition unlimited and unconstrained; to place a limit on a certain freedom, then, seems paradoxical. The process of considering limits on freedom of expression demands that we also consider what expression is, what functions it serves in society, why it is valuable, and to what extent this value protects it from limitation. We must confront the fact that expression does not exist in a vacuum, but as a basic part of human life which is inherently connected to political ideas regarding how society should be organized. Therefore, it is extremely difficult, or perhaps impossible, to make objective decisions about what falls under the umbrella of "free expression". This is true even if we follow the liberal assumption that, in order to preserve individual freedom, it is only appropriate to determine what society *ought not* to look like, rather than forming a positive image of what shape it *ought* to take. The harm principle is a case in point of this approach, in that it aims to determine outcomes to be avoided, rather than those to be desired. Still, this requires an objective judgment of what is undesirable in society; the negative, in some sense, implies the positive as well. In the following sections, I will provide a brief overview of some of the common theoretical issues regarding the application and feasibility of the harm principle.

a. Issues of breadth and subjectivity

Mill's position concerning harm, which can be taken to represent the standard liberal position on the limits of individual liberty, is that society may only limit the acts of individuals when those acts cause harm to others. This conclusion follows from the fact that individuals, although free to act as autonomous beings, never live in isolation, but rather in a human community of constant influence, interaction, and exchange. Human relationships, however, are nuanced and resistant to generalization. An exchange between two individuals may mean something different depending on who is involved, the context of the exchange, and an indeterminate number of other factors. This point, which may seem inappropriately broad, suggests that there is no way to standardize “harm”—no way to determine with certainty that a particular act will always be harmful, no way to integrate or harmonize different opinions about what constitutes harm. As a result of this difficulty, the interpretation and applicability of the harm principle remains a contested issue within legal and political scholarship.

David Lyon writes about the difficulties of the openness of Mill’s harm principle, making a simple yet transformational observation: “the principle of liberty is narrower than a principle of utility in two ways: it concerns harm to others, not welfare generally…”[13] While the distinction between harm and general welfare does not straightforwardly solve the problem of what Mill may have meant by “harm”, it does help in the process of defining harm. It suggests that the prevention of harm does not necessarily equal the promotion of welfare, which is to say that individual liberties may not be limited in the interest of promoting welfare unless it is also shown that a lack of interference will result in harm. This seems to narrow the definition of harm. Lyon also considers whether the *failure* to act in some cases may also constitute harm; he claims that “in Mill’s view, one may legitimately be required… to cooperate in joint undertakings and to act as a good samaritan.”[14] Here, the definition of harm appears much broader. He ultimately proposes that:

> freedom may be limited only for the purpose of preventing harm to others, but the conduct that is interfered with need not be considered harmful or dangerous to others… The cooperation and good samaritan requirements that Mill refers to could not be justified on the ground that they prevent conduct that causes harm to others; but it can be argued that such regulations nevertheless work in other ways to prevent harm to others.[15]

While the final question of what constitutes harm is not the focus of the present discussion, it is still useful to consider this definitional problem in order to understand that the term “harm” is somewhat subjective.

Dudley Knowles reflects on the issue of defining harm: “what Mill does not provide in *On Liberty*, and what his argument demands, is a clear conception of direct harm.”[16] The ambiguity of what constitutes harm in liberal theory follows from the subjective nature of the

[13] David Lyons, “Liberty and Harm to Others,” in *Mill’s On Liberty: Critical Essays*, edited by Gerald Dworkin. (Maryland: Rowman & Littlefield Publishers, Inc., 1997): 115. Here, Lyons refers to the harm principle rather as the “principle of liberty”. This choice of language, where the harm principle essentially or characteristically stands for the theory of liberal-*ism*, is influential in my assumption of the harm principle as a focal point.

[14] Lyons, “Liberty”.

[15] Lyons, “Liberty,” 118.

[16] Dudley R. Knowles, “A Reformulation of the Harm Principle,” *Political Theory* 6, no. 2 (May 1978): 235.

concept of "harm" in language more generally. Not only is there the issue of scale—minor inconveniences and major catastrophes can both be called harmful outcomes—but also the issue of how "harm" is intrinsically connected to morality. What I mean to suggest is that the notion of "harm" is based on the existence of moral good. This is not to say that all frameworks of moral good are appropriate to apply to society at large—this is, on the contrary, what liberalism seeks to avoid—but that it is shortsighted to imagine a liberal social order which completely omits concerns of "final ends". In order to be able to maintain a liberal approach in different social and cultural circumstances, it is necessary to acknowledge that even the values of liberalism have some moral content, or at least cannot be absolutely separated from moral concerns.

A final relevant issue with the harm principle is the impossibility of determining the harmfulness of individual acts. Knowles points to this problem: "for any action description, it is possible that some actions that satisfy that description cause harm to others."[17] What follows is that any act is potentially harmful, and that the determination of whether an act has caused harm will depend on an evaluation of the context of its occurrence. While "legislation is standardly justified in terms of generalizations assigning some probability of harm to actions which exemplify the practice in question," such generalizations cannot paint a representative picture of each and every circumstance.[18] The significance of an act can only be determined on an individual, particular level; this, however, is clearly not an option when it comes to establishing law. Of course, the practices of legal administration are designed to respond to individual cases; this is the role of the judge. However, it must be acknowledged that the greater the discretion of the judge in evaluating and deciding on invidual cases, the greater the room for the application of subjective judgments, which is to be minimized.

As a result of these several issues, restrictions of certain acts are not entirely separable from moral concerns; in this way, the harm principle sits uneasily with the notion of liberalism as a theory which avoids governance by moral standards.

b. Alternative approaches to the harm principle: restrictive versus expansive readings

If the harm principle is overly broad, it may fail to protect individual liberty in any meaningful way. The concern, in short, is that "if 'harm' is not severely restricted to something

[17] Knowles, "A Reformulation," 239.

[18] Knowles, "A Reformulation," 241.

like a rights violation, then the harm principle cannot do the work it is intended to do."[19] The threshold of harm as an act that threatens someone else's rights is a common restriction of the harm principle. Joel Feinberg proposes such a reading of the harm principle, going so far as to propose a separate *offense* principle which applies to acts where an individual does wrong but perhaps does not pass the threshold of harm. He explains that "even in the possible cases where the threshold of actual harm had not been reached, a serious wrong was surely done them," and suggests that these wrongs must also be legally preventable in some cases.[20] By this view, an offensive act is harmful in a secondary way, while a harmful act must be harmful in itself. Feinberg seems to understand that traditional liberalism does not protect the individual from the moral harm of offensive acts: "even though the protection of moral sensibility from profound offense does not logically imply the protection of persons from *any* kind of distress, it is true as a matter of empirical fact (and this was Mill's major emphasis) that legislatures are prone to slide in that direction once they start down the slope."[21] Again, we are reminded of the oppressive tendency of majoritarian moral thinking which was an original concern of liberal theory.

On the other hand, an expansive reading of the harm principle is also possible; this is the approach that Piers Turner suggests in his writing. He claims that restricted conceptions of "harm" face serious interpretive challenges, and that the harm principle is best understood as simply an anti-paternalism principle.[22] This perspective is particularly important to further my argument that there is no way to completely exclude moral concerns from the process of regulating human action, and that Mill's theory of liberty should not necessarily be interpreted as the total abolition of legal moralism from governance. Turner writes that "Given the ambiguity surrounding 'harm' in On Liberty itself, it is instructive that a survey of Mill's Collected Works reveals 'harm' to be consistently used in practical contexts as simply a companion term for 'good,' just as 'cost' is now a companion to 'benefit.'"[23] This view is consistent with the notion that harm can only accurately be determined by putting acts into context on an individual case basis. It reformulates the harm principle such that "'harm to others' includes any bad consequence for others, and the purpose of the harm principle is only

[19] Turner, "Harm," 303.

[20] Joel Feinberg, "Profound Offense," in *Mill's On Liberty: Critical Essays*, edited by Gerald Dworkin (Maryland: Rowman & Littlefield Publishers, Inc., 1997): 139.

[21] Feinberg, "Profound Offense," 155.

[22] Turner, "Harm," 319, 325.

[23] Turner, "Harm," 320.

to exclude paternalistic considerations from social deliberation."[24] It is my view that Turner's position lends us a more realistic view of the place of social morals in the process of determining harm and therefore the limits of individual liberty.

c. Scanlon's theory of freedom of expression

Philosopher T.M. Scanlon, writing from the American perspective, proposes a theory of freedom of expression which deals (albeit indirectly) with the difficulties of the harm principle I have raised so far.[25] His argument centers on responding to claims of the "irrationality" of the doctrine of freedom of expression; that it establishes classes of acts which are protected from government interference, "despite the fact that they have as consequences harms which would normally be sufficient to justify the imposition of legal sanctions."[26] Scanlon's observation here, that some harms are seen to warrant legal prohibition of the inciting act while others are perhaps not "serious" enough to justify limitation of an act, reveals that free expression doctrine partially depends on a balancing of good and bad consequences. There are, in some sense, two competing harms: the harm of the act itself, and the potential harm done to society by limiting the speech without precise necessity. As a result, some harmful acts are in a sense not harmful *enough* to call for limitation. This approach to the doctrine of freedom of expression is known as the "consequentialist" approach.

Scanlon almost entirely dismisses the consequentialist approach; his commentary suggests that he believes there is no way to accurately generalize the harm caused by certain acts when considered categorically as "expression". He explains, for example, that "since acts of expression can be both violent and arbitrarily destructive, it seems unlikely that anyone would maintain that as a class they were immune from legal restrictions. Thus the class of protected acts must be some proper subset of this class."[27] Scanlon claims that such an approach is "clearly wrong", as it is prohibitively difficult to distinguish between protected and unprotected acts.[28] As a result, Scanlon follows an alternative approach: "Rather than trying at

[24] Turner, "Harm," 322.

[25] What I mean by the "American perspective" here is the approach to freedom of expression whereby anything considered to be "speech" is protected, and where therefore the theoretical issue lies in determining whether certain acts constitute speech. This is in contrast to the perspective where speech may be limited if a process of balancing interests demonstrates that limitation would be justified. Such a perspective characterizes the European approach to freedom of expression.

[26] Thomas Scanlon, "A Theory of Freedom of Expression," *Philosophy & Public Affairs* 1, No. 2 (Winter 1972): 204.

[27] Scanlon, "A Theory of Freedom of Expression," 207.
[28] Scanlon, "A Theory of Freedom of Expression," 208.

the outset to carve out the privileged subset of acts of expression, then, I propose to consider the class as a whole and to look for ways in which the charge of irrationality brought against the doctrine of freedom of expression might be answered without reference to a single class of privileged acts."[29]

If we cannot define the protected class of speech, then we cannot evaluate the harm of an instance of expression by fitting it into a category of a certain type of speech.

There are multiple aspects of Scanlon's approach to freedom of expression; my focus will be on what he calls "the Millian Principle", which deals with the issue of harm. This principle constructs the freedom of expression by naming certain harms that may *not* be used to justify limitation of expression; it is, in this way, a negative construction of what may constitute freedom of expression. The principle is as follows:

> There are certain harms which, although they would not occur but for certain acts of expression, nonetheless cannot be taken as part of a justification for legal restrictions on these acts. These harms are: (a) harms to certain individuals which consist in their coming to have false beliefs as a result of those acts of expression; (b) harmful consequences of acts performed as a result of those acts of expression, where the connection between the acts of expression and the subsequent harmful acts consists merely in the fact that the act of expression led the agents to believe (or increased their tendency to believe) these acts to be worth performing.[30]

Rather than create a *threshold* of harm, this principle sets out certain harms which are exempted from Mill's harm principle. These harms, notably, are not exempted because of the severity or extent of their impact *per se,* but rather because of the risk involved in allowing the state to identify them as harmful. Scanlon explains that:

> for a law to provide such protection [against the harm of coming to have false beliefs] it would have to be in effect and deterring potential misleaders while the potentially misled remained susceptible to persuasion by them. In order to be protected by such a law a person would thus have to concede to the state the right to decide that certain views were false and, once it had so decided, to prevent him from hearing them.[31]

The notion that the state should not have the right to decide that certain views are false is in line with Mill's liberal understanding of the relationship between society and the individual. Scanlon suggests that this may be called "the right of citizens to make up their own minds"—if the "Millian Principle" were seen to rest on a right.[32] This aspect of Scanlon's theory will be relevant in subsequent sections of this piece which concern the regulation of misinformation.

[29] Scanlon, "A Theory of Freedom of Expression," 208.
[30] Scanlon, "A Theory of Freedom of Expression," 213.
[31] Scanlon, "A Theory of Freedom of Expression," 217.
[32] Scanlon, "A Theory of Freedom of Expression," 221.

III. Conclusion- moral concerns, truth, democracy

Naturally, with its emphasis on preserving social pluralism, a liberal democracy (as influenced by traditional liberal theory) positions itself as morally neutral. After all, a primary aim of the liberal approach is to prevent the oppressive dominance of popular opinion. Liberalism is an ideology centered around the concepts of the liberty and equality of the individual in the public sphere. Importantly, these concepts are also values, and they are values which are not completely separate from morality. Although the political values of liberty and equality are admittedly not comparable to religious moral values which have often bled into systems of governance and politics, it would be an oversight to dismiss the connection between the two. The basic, structuring values of a political philosophy are functionally similar to more "obvious" moral values in that they both propose a certain vision of how human life and society should be organized. I believe that acknowledging this functional similarity helps us to understand the liberal defense of free expression and to clarify the harm principle. It takes the discussion of harm from a highly broad and theoretical level back into the world of real experience, where structural principles (like morality and political values) are essential. Put another way, an act which deeply offends or attacks the political values of liberalism, which become integrated with individual rights, may also be seen as harm which calls for interference. I mention this interpretation in the context of expression because acts of expression and speech are particularly likely to pose ambiguous or non-physical harm. The harm in discriminatory or hateful speech, for example, is altogether different from the harm of physical attack or interference with a person's body or property, which is different from the potential harm of misinformation and false claims of fact. The rapid and mass nature of digital expression means that we now communicate, learn, and form judgments differently than in the pre-digital era. It therefore has different potential effects, and different potential harms—many of which will be of such an ambiguous or non-physical kind. When attacks on political values or tenets—attacks which *destabilize concepts* rather than directly impact bodies or property—are considered as harms, we must also revisit the meaning and interpretation of such concepts. I will argue that the contemporary world of expression, crucially shaped by digital media, calls for a re-evaluation of how the values of liberal democracy are best upheld in the realm of expression.

Chapter Two: Misinformation in the digital era: determining harm, regulating speech

I. The politics of factual truth

The problem of misinformation has strongly reared its head in recent years, as the digital world expands and so many aspects of our lives, both personal and political (private and public), are shaped by the flow of information on the internet. The terms *misinformation, disinformation*, and *fake news* are now very frequently circulated in the world of politics and public affairs. While each term holds distinct connotations—think of the accusations from the American right of the "fake news" media—the popular usage of these terms points more generally to the instability of "truth" in the contemporary world of information. Similarly, it is highly popular for social and political commentaries to lament over the ushering in of a "post-truth" era.[33] I do not, however, plan to remark on the current circumstances of "post-truth" politics, which would require a dive into the political rhetoric and narratives of right-wing populists. Rather, I make mention of this situation, of truth's position as a contentious political item, because it shares a common vocabulary with the issue at hand, that is, the regulation of misinformation. The regulation of misinformation on digital platforms is plainly an issue of freedom of expression, and therefore is inherently political. It would be a mistake to omit mention of the panic surrounding the concept of truth in recent politics, as this and digital misinformation are concurrent phenomena. In other words, at the same time that a number of liberal democracies are coping with the populist political trauma of managing a "post-truth" society, the reality of misinformation in the digital world brings related questions of truth and authority onto the table.

I also do not mean to suggest that the crisis of truth in politics has emerged only in the contemporary world; in fact, the opposite is true, and the long-standing nature of the tension between the two is indeed theoretically relevant in the present discussion on approaching the issue of misinformation as basically *un*-truthful speech. In her notable *New Yorker* essay "Truth and Politics", Hannah Arendt begins by observing that "no one has ever doubted that truth and politics are on rather bad terms with each other, and no one, as far as I know, has ever counted truthfulness among the political virtues."[34] From this perspective, the problematic relationship between truth and politics means that lies are nearly always expected to be a piece woven into

[33] *Post-truth* was the Oxford Dictionaries 2016 word of the year; It is defined as an adjective relating to circumstances in which objective facts are less influential in shaping public opinion than emotional appeals.

[34] Hannah Arendt, "Truth and Politics," *The New Yorker*, 25 February, 1967.

the fabric of political narratives. We should, however, refrain from passively accepting this situation; in fact, in order to understand what is at stake in approaching the topic of misinformation, we should return to the question of the relationship between truth and politics, and more particularly within the context of liberal democracy. To this point, Arendt writes that "while probably no former time tolerated so many diverse opinions on religious or philosophical matters, factual truth, if it happens to oppose a given group's profit or pleasure, is greeted today with greater hostility than ever before."[35] This brings us back to the notion that the aim of liberalism was to take "final ends" off of the table, in the sense of making concrete the divide between church and state. In a political system where religion and philosophy cannot serve as frameworks of definitive truth to structure society, indeed *any* claim to a "final end", including claims to factual truth, will encounter resistance. They easily appear as attempts to grab power by suppressing dialogue and stubbornly proclaiming the end of political discussion. What Arendt means by truth and politics being "on rather bad terms" is that truthfulness, as in factual truth, is not in and of itself considered politically valuable. Whether or not this is really the case, it is apparent that the politics of liberal democracies *do* depend on factual truth.

In a system where the individual's political judgments are (at least in theory) the basis for political and government action, factual truth serves as the departure point for deliberation and opinion formation. By this understanding, the problem of misinformation is easily solved: claims of factual truth can, and ought to be, regulated by trusted authorities (traditionally, this would mean government; but in the digital era, this means social media companies). The rationale for allowing such entities to monitor claims of factual truth involves notions of harm prevention and public interest concerns; in other words, the nature of the authority's responsibility in governance means that it has an incentive to censor false claims of fact that it believes to pose a risk of harm to the society. This position also assumes the impartiality, and indeed infallibility, of the authority on issues of public interest and on determinations of truth. However, the blurred dividing lines between fact, opinion, and interpretation make the project of regulating misinformation incredibly difficult, if the standards of freedom of expression are to be preserved. There is a tension between the potential harm of false claims of fact and the illiberal move of censoring opinions and interpretations. Arendt raises this point:

> But do facts, independent of opinion and interpretation, exist at all? Have not generations of historians and philosophers of history demonstrated the impossibility of ascertaining facts without interpretation, since they must first be picked out of a chaos

[35] Arendt, "Truth and Politics".

> of sheer happenings… and then be fitted into a story that can be told only in a certain perspective, which has nothing to do with the original occurrence?[36]

The difficulty of distinguishing between what is a claim of fact and what is an opinion or interpretation of fact is heightened by the sheer scale of speech on social media; how can such a process of distinction possibly be standardizable or able to be integrated into speech platforms via algorithm alone? If matters of opinion are accidentally taken as false claims of factual truth in the process of content regulation on a basis of "truthfulness", there is the risk of suppressing democratic political life. It goes without saying that this potentiality does not sit well with the liberal approach to freedom of expression. The next question will concern the potential harms of misinformation, and whether these harms are more significant than the risk of unduly constricting politics through the regulation of speech.

II. The nature of the social media company's power

As a private actor, Facebook holds the right to regulate content as it sees fit, and can impose content restrictions without legally violating its users' right to freedom of expression. Following this, one can argue that it is not appropriate to evaluate social media content regulation policies as if they were freedom of expression provisions. From this view, the social media company's ability to regulate content may be compared to the editorial privilege that publishers enjoy in deciding what content they will publish; this comparison does not, however, imply that social media companies hold traditional publishing liability—importantly, they do not.[37] Still, such a comparison is conceptually useful, as it draws the conclusion that the right to freedom of expression is ultimately unaffected by the behavior of private actors. Although this may be true in a strictly technical sense, it must also be taken into consideration that social media companies play a powerful and constitutive role in the public sphere, such that their actions regarding speech regulation *do*, in fact, have a similar effect to state actions. As the platforms for an increasingly large proportion of all speech and expression, social media companies have an unprecedented influence on the status of freedom of expression, despite the fact that the right to freedom of expression itself remains constant. Professor Kyle Langvardt goes even further, arguing that "those [social media] corporations' power over public discourse today is so concentrated and far- reaching that it resembles and *arguably surpasses state power* within its sphere [emphasis added]."[38]

[36] Arendt, "Truth and Politics".

[37] This is an imperfect comparison, as social media companies are platforms or intermediaries, rather than speakers.

[38] Kyle Langvardt, "Regulating Online Content Moderation," *The Georgetown Law Journal* 106 (2018): 1357.

Facebook's Community Standards (hereafter the "Standards" or "Community Standards" in some cases) have the expressed goal of "[creating] a place for expression and [giving] people a voice."[39] Here, the language of "creating" and "giving" reveal the company's awareness of its power in the realm of speech in the digital era; an actor that has the power to create a space of expression and give people a voice also has the power to limit the size of that space and to decide the volume of people's voices. Once a company takes on the task of providing a platform for speech, it also adopts, in some sense, the responsibility for regulating that speech (if only in the interest of its own financial concerns, or in the interest of its positive public image). Of course, a digital platform company may also have its own political and social biases which influence this sense of responsibility—it may aim, for example, to win over libertarians or those on the far-right who prize unmoderated digital spaces.[40] Despite the public nature of digital expression, its regulation is still private; this poses definite risks to freedom of expression, as legal standards for speech do not have binding force on private companies. Imanol Ramírez presents the current challenge effectively: "Just as there is a critical need of moderation [due to the large volumes of potentially harmful content], the danger of censorship is amplified online because online intermediaries control a vast share of communications while maintaining the power to mediate these communications."[41] The core of the issue is that, due to the private-actor nature of social media companies, we lack a framework for their legal accountability.

The perspective which recognizes the similarity between private and public power in the realm of expression is the one which I will follow for the remainder of my writing here. The understanding that the nature of the social media company's power in society is somewhat parallel to that of government power in regulating speech is the grounding justification for my analysis of Facebook policy using the terms of liberal political theory; in discussing the challenges of regulating digital speech according to liberal principles, I am to suggest that Facebook may in fact be exceeding the authority that liberal speech theory would lend it as a regulatory power. Clearly, the advent of the digital age has changed the nature of power in society and government; this shift in the balance of power between private and public actors

[39] "Facebook Community Standards." *Facebook, Inc.*, 2021. https://www.facebook.com/communitystandards/.

[40] This was the case with Parler, a short-lived online platform that attracted fringe political groups and was taken offline after some of its users participated in the storming of the U.S. Capitol on January 6, 2021.

[41] Imanol Ramírez. "Online Content Regulation and Competition Policy." *Harvard Law School Antitrust* Association, December 3, 2020. https://orgs.law.harvard.edu/antitrust/files/2020/12/Imanol-Ramirez-Online-Content-Regulation-and-Competition-Policy-HLSAntitrustBlog-2020.pdf

also straightforwardly warrants a return to the grounding political theories which have informed our ideas of power and governance. Langvardt points to the legal perspective of this shift: "if nothing else, the twentieth century's law of free expression established that only the final censor—at that time, the state—was subject to law. Today, a small number of politically unaccountable technology oligarchs exercise state-like censorship powers without any similar limitation."[42] Therefore, in my view, it makes sense to evaluate the actions of these incredibly powerful companies against the standards of political theory which have shaped past government policies on freedom of expression. Liberal and democratic political theory provides a framework for thinking about the power relationship between the individual and large, influential companies like Facebook just as it does for the relationship between the individual and their government.

a. Free speech on Facebook, in its own words

Facebook is a large and influential private actor which, despite its state-like power, is not held to the same standards of legality and accountability as governments. Notably, although Facebook and other social media companies have a state-like power in the realm of deciding what speech is present in the public sphere, their power to punish those who break the rules is entirely different—Facebook itself cannot impose a legal punishment. While the company is not bound to the standards that a government must meet, it would be a mischaracterization to say that Facebook has wholly dismissed the responsibilities that it incurs as a result of its power in society. The nature of large social media companies' power is a question unique to the present moment of neoliberal capitalism, and the appropriateness of how these companies have chosen to organize and "check" their own power is outside the scope of the present line of inquiry. Still, despite this lack of clarity regarding the future of big tech and government, an analysis of how these companies imagine their relationship to current legal standards is crucial in thinking about the future of individual rights like freedom of expression. Of course, we must also keep in mind that an expressed commitment to meeting certain legal standards may not be faithfully executed, for a number of reasons.[43]

[42] Langvardt, "Regulating Online Content Moderation," 1358.

[43] Again, the scale of content poses an issue for content moderation. The inability of algorithms to make sufficiently nuanced judgments about content shows the need for human involvement in the process; however, even with human content moderators, issues of scale and context reappear. Kyle Langvardt, cited above, refers to the "erratic and opaque decision-making" involved in Facebook's content moderation. Similarly, a report from NPR notes that much of Facebook's content moderation is carried out by overworked subcontractors, who often have to make decisions on the basis of little to no context. In the report, a former employee of Facebook explains that "the subcontractor 'pretty much tosses a coin'" in determining whether or not a post violates standards or not.

A 2018 article by Richard Allan, Facebook's Vice President of Policy, addresses the platform's approach to free expression. The article is part of a series from Facebook called *Hard Questions* which "addresses the impact of our products on society." In it, Allan explicitly mentions that Facebook is not a government, and that they aim to moderate content "in a way that gives free expression maximum possible range." He notes that there are "critical exceptions" to freedom of expression on Facebook, directing the reader to the Community Standards. Importantly, he explains that while Facebook is not bound by international human rights law, they are "part of a global initiative that offers internet companies a framework for applying human rights principles to our platforms," and that they look towards such legal standards in the area of freedom of expression.[44] Unsurprisingly, the harm principle emerges as a core concept of the legal standards of freedom of expression that Facebook adopts in its approach on drawing the line of free expression.[45]

In the *Hard Questions* article mentioned above, Allan takes the leap into defining what "harm" means on Facebook. The categories included are: 1) personal harm (speech that poses a credible threat of violence) and 2) hate speech.[46] The discussion of defining harm anticipates the issue of the harm in misinformation; it mentions that "whether or not a Facebook post is accurate is not itself a reason to block it." This is to say that the individual has a right to say things that are not true; in explaining why Facebook has adopted this policy, Allen again cites the standards of human rights law (though without mention of a specific provision or concept). He also mentions an approach which I will address in greater detail later; that is, that "rather than blocking content for being untrue, we [Facebook] demote posts in the News Feed when rated false by fact-checkers and also point people to accurate articles on the same subject."[47] From the view presented here, misinformation generally does not constitute a harm great enough to warrant the limitation of speech. It remains to be seen, however, whether or not the approach taken in this area is consistent with the traditional liberal conception of free expression.

https://www.npr.org/sections/alltechconsidered/2016/11/17/495827410/from-hate-speech-to-fake-news-the-content-crisis-facing-mark-zuckerberg

[44] Richard Allan, "Hard Questions: Where Do We Draw the Line on Free Expression?" *Facebook, Inc.*, August 9, 2018. https://about.fb.com/news/2018/08/hard-questions-free-expression/

[45] International Covenant on Civil and Political Rights (ICCPR) Article 19. Article 19 says that the right to freedom of expression may only be limited when they are "provided by law and are necessary for (a) For respect of the rights or reputations of others; or (b) For the protection of national security or of public order (ordre public), or of public health or morals."

[46] In 2017, Allan wrote another piece in the *Hard Questions* series which addresses "Who Should Decide What Is Hate Speech in an Online Global Community?"

[47] Allan, "Hard Questions." https://about.fb.com/news/2018/08/hard-questions-free-expression/

III. COVID-19 misinformation on Facebook

The contemporary context of the COVID-19 pandemic is exemplary of a situation where the dividing line between opinion and factual claim is blurred. It is also a situation where the place of this dividing line is particularly consequential; regardless of the incalculable disputes over facts, it remains that there is, at the ground, the reality of a public health crisis that must be managed. In other words, there are certain factual realities which we must respond to. In Arendt's phrasing, "conceptually, we may call truth what we cannot change; metaphorically, it is the ground on which we stand and the sky that stretches above us."[48] The challenge in approaching misinformation lies partially in the difficulty of identifying what we believe to be the basic factual conditions of the situation at hand. These conditions, once identified, are then 1) officially recognized, and therefore lent an image of validity and truthfulness, and 2) considered harmful if replaced by lies. This decision is consequential in determining the shape of the sphere of public discourse.

Social media companies, as large platforms for public speech in the digital world, have already come up against this challenge; in the next sections, I will focus more specifically on Facebook's response to pandemic-related misinformation in 2020. My initial focus will be directed towards the approach of the Community Standards regarding the concept of "harm" in general, and then within the context of misinformation. I will include a detailed analysis of the Facebook Oversight Board's decision 2020-006-FB-FBR, which concerns Facebook's decision to remove a post which contained alleged misinformation on COVID-19 and was seen to violate a policy of Facebook's Community Standards.

a. Harm and misinformation in the Facebook Community Standards

In line with the assumption that Facebook's power in regulating speech is not unlike the power of the state in doing so, the Community Standards begin with the explanation that "we [Facebook] recognize how important it is for Facebook to be a place where people feel empowered to communicate, and we take seriously our role in keeping abuse off our service."[49] This organizational intention expresses both an acknowledgment of the importance of freedom of expression (or at least the freedom to communicate) and of the risks of harm inherent in expression. It recognizes that there is a point at which speech can, and perhaps must, be limited. While *abuse* and *harm* are not equal terms, to speak of the priority of avoiding abuse in a community, as the Community Standards do, is similar to speaking of the necessity of avoiding

[48] Arendt, "Truth and Politics".

[49] "Facebook Community Standards"

harm to others. This reference to the harm principle of liberal theory is one way in which the Community Standards can be seen to reach towards liberal standards of freedom of expression, in the sense that Facebook limits its regulatory authority to harmful content.

Of course, harm is mentioned more explicitly later in the Community Standards. For example, Part 1 of the Standards, which focuses on violence and criminal behavior, includes the "aim to prevent potential offline harm that may be related to content on Facebook." Here, the specific mention of *offline* harm implies that the content or speech itself is not harmful, but rather that its potential consequences may be, and that that potentiality is sufficient to warrant limitation by the platform. According to the Standards, however, the decision to limit potentially harmful content is made following a different process depending on the *type* of content. In this way, Facebook recognizes two categories of potentially harmful content: that content which may be regulated without additional context or justification, and that content which Facebook requires "additional information and/or context to enforce." This distinction implies that content falling into the latter category is content which is not harmful in itself, but which may be part of a greater context in which it poses a direct risk of harm. Within the Community Standards, misinformation falls into this second category: "misinformation and unverifiable rumors that contribute to the risk of imminent violence or physical harm."[50] As I will lay out below, this is the relevant standard in the Oversight Board's decision concerning COVID-19 misinformation.

b. Case overview: Oversight Board decision on COVID-19 misinformation

The Facebook Oversight Board's decision concern's Facebook's removal of a post made in October 2020 which was designated as health misinformation that "contributes to the risk of imminent physical harm". The post was made in a public Facebook group related to COVID-19, and included 1) an accusation of poor health strategy by the French government and 2) criticism of a government agency's decision to refuse one treatment but approve another; this latter criticism made mention of the refused treatment as a "cure" advocated by Didier Raoult, a professor at a French medical university. The group where the post was shared had more than 500,000 members, which is relatively small when seen against the billions of Facebook users globally, but is on the other hand a considerably large audience for an act of public speech made by an ordinary individual. In any case, the post was seen to pose a genuine risk of harm under the Community Standard on Violence and Incitement, and was removed in

[50] "Facebook Community Standards"

the interest of preventing offline harm. The issue before the Oversight Board, then, was whether the removal of the post complied with the Community Standards, Facebook's values, and international human rights standards; the Board ruled in the negative, and required that the post be restored.

In its case decision, the Oversight Board refers to the relevant standard of the case as the "misinformation and imminent harm rule"; I will do the same here. The Board's analysis of Facebook's compliance with the Community Standards in removing the post in question includes the explanation as to why the post was considered as misinformation and in what way it was considered to be harmful. The following is provided:

> Facebook stated the post constituted misinformation before it asserted there was a cure for COVID-19 whereas the WHO and leading health experts had found there is no cure. Facebook noted that leading experts had advised the platform that COVID-19 misinformation can be harmful because, if those reading misinformation believe it, then they may disregard precautionary health guidance and/or self-medicate. Facebook relied on this general expert advice to assert that the post in question could contribute to imminent physical harm.[51]

Ultimately, the Board holds that Facebook did not provide sufficient evidence that the post meets its own standard on imminent harm, and that therefore it had violated its own Community Standard. More specifically, the Board explains that there was no real risk of harm involved, as the proposed "cure" to COVID-19 was not made more available to the public via speech that advocates for its use. The case decision includes a very brief comment (two sentences) on Facebook's compliance with its own values; it concludes that the rationale in removing the post did not properly balance the values of "Safety" and "Voice". Using the language of harm and freedom of expression, this is to say that the harmfulness of the post was overestimated, and that the jurisdiction of the platform to limit the speech was invoked without sufficient justification.

The final compliance analysis that the Oversight Board provides in the case concerns Facebook's compliance with human rights standards on freedom of expression. The Board notes that, according to the United Nations Guiding Principles on Business and Human Rights (UNGPs), companies are expected to comply with the human rights standards of international law and to be aware of and takes steps to address the ways in which their work may have negative consequences for human rights.[52] Although Facebook is not technically bound to legal standards of freedom of expression, the Oversight Board's application of the standards of

[51] *Case Decision 2020-006-FB-FBR.* Facebook Oversight Board (2020). https://oversightboard.com/decision/FB-XWJQBU9A/

[52] Ibid.

international human rights law to the case provides an insight into both the case at hand regarding health misinformation *and* into the general difficulties of integrating typically liberal democratic standards of freedom of expression into the regulation frameworks of digital spaces, and especially those difficulties surrounding the treatment of online misinformation.

The Board measures Facebook's moderation decision against the freedom of expression standards of Article 19 of the ICCPR and provides that any restriction on this freedom should meet the conditions of legality, legitimacy, and necessity. The first part of the test, which evaluates the legality of the measure taken, in this case "requires assessing whether the misinformation and imminent harm rule is inappropriately vague."[53]

The second part of the test concerns whether Facebook had a legitimate aim in removing the post. Here, the Board quickly concludes that Facebook satisfied the test by specifying the aim of protecting public health during a global pandemic.[54] Finally, the third part of the test requires that Facebook demonstrate that removal of the post was the "least intrusive means" possible to achieve the aim of protecting public health.[55] The Board notes that Facebook was unable to provide a proper explanation as to why removal was the least intrusive means, and therefore holds that the action fails the necessity test. It also cites how the Community Standard on False news provides several alternative means of responding to misinformation, including:

> the disruption of economic incentives for people and pages that promote misinformation; the reduction of the distribution of content rated false by independent fact checkers; and the ability to counter misinformation by providing users with additional context and information about a particular post, including through Facebook's COVID-19 Information Center.[56]

Since the regulation of misinformation online is a nearly brand-new task for companies (and perhaps soon for legislators), the impact of these various means of response on freedom of expression remains to be evaluated. A basic assessment of these means of speech regulation in the context of freedom of expression principles will follow in the final sections of this paper.

[53] Ibid.
[54] Ibid.
[55] Ibid.
[56] Ibid.

Conclusion: Challenges of a liberal response to misinformation & the path forward

The work of the Oversight Board on an isolated instance of misinformation regulation reveals how much work is ahead in the future of online expression more broadly. The Board's compliance analysis in the case here is in many ways inconclusive, appearing as a temporary solution to a much larger problem, as it raises extremely complex issues, including the evaluation of the harm of online speech, the difficulty of distinguishing between opinion and fact, and the possibility of alternative means of limiting (but not removing) speech on social media. Each of these issues is theoretically interesting and consequential for the concept of freedom of expression. I will briefly address these theoretical issues to suggest that the feasibility of maintaining a liberal approach to freedom of expression in the expanding digital world is decreasing, especially as we face the time-sensitive problem of misinformation (in all forms—political, social, economic, etc). The future of online speech regulation is tenuous, whether it is taken up by governments or left to private companies.

The perennial issue of determining the harm of expression has been significant for Facebook in the development and execution of its content moderation policies. Despite the company's efforts to define harm such that harmful content is easily recognizable and therefore easily regulated, the lack of clarity of what "harmful content" looks like in practice suggests that harm may not be a workable standard for the limitation of online expression. The first challenge that social media companies face in determining standards of harm is identifying the causal relationship between online expression and offline harm. In its Community Standards, Facebook identifies a number of absolutely prohibited content, where the connection to harm is self-apparent (such as in the case of calls to violence or organization of crime); however, the reality is that not all potentially harmful speech will fit into these boxes. This results in a situation where regulation will be either chronically over- or under- restrictive. The challenge of determining the real-world effects of digital expression also makes it difficult for platforms to adhere to the liberal standard of limiting only *direct* harms, as traditional liberal theory would require. Additionally, as seen in standard addressed the Oversight Board case above, Facebook integrates language of *contribution* to the risk of harm, to the effect that content need only contribute to the risk of harm, rather than directly cause it; this clearly contradicts the liberal standard of direct harm.

The massive scale of online expression also poses a challenge for companies seeking to adhere to the liberal harm principle. This is because, as discussed earlier, harm is not only a subjective concept, but it is also heavily context dependent.[57] This is to say that the same speech may mean something different, carrying different potential harms, depending on the context of its expression. The tone-deafness of most digital, text or image-only speech does not ease the matter. It goes without saying that a context-specific evaluation of harm cannot, with the current level of technology, be standardized to the extent that digital platforms would require; neither an algorithm nor teams of human content moderators have the capacity to make these nuanced judgments at the necessary speed. Whether this is a problem to be resolved by technological solutions or by an increase in the staffing of social media companies remains to be seen.

If we pivot and think of the harm principle as simply an anti-paternalism principle for digital expression, as suggested by Piers Turner, we run into similarly prohibitive difficulties.[58] This is especially true for the regulation of misinformation; it may be said that the limitation of misinformation is an inherently paternalistic act, as it denies the individual the right to reach their own conclusions. Indeed, the notion that false claims of fact should be removed from the public sphere of expression contradicts Mill's concept of the "marketplace of ideas", where it is accepted that all opinions may contain partial truths, and that the falsity of speech may not justify its limitation or censorship, given that there is adequate space for counter-speech. Here, we can also return to T.M. Scanlon's "Millian principle", which implicitly rests on "the right of citizens to make up their own minds."[59] From this perspective, misinformation cannot be said to be harmful enough to justify limitation. It is fundamentally an anti-paternalist principle, which assumes the equal capacity for individual judgment among all individuals, and which would lead to the claim that any harm caused by misinformation is only the business of decision-makers, and not a governing authority. In light of contemporary issues of misinformation online, where public conversations, and indeed the policymakers of social media companies themselves, overwhelmingly ignore the possibility of taking a completely hands-off regulatory approach, this seems like an extreme position. The numerous controversies over misinformation, seen on both the political right and left, are cases in point of the existence of a strong will to respond to misinformation through policy; they also indicate the extent to which the topic of misinformation has been politicized.[60] However, as seen here,

[57] In the section "Issues of breadth and subjectivity" of the present work, see previous comments about the work of Dudley Knowles on the harm principle.
[58] See the section "Alternative approaches…" of the present work.
[59] See the section "Scanlon's theory of freedom of expression" of the present work.
[60] Again, this relates to the crisis over truth and information that has surged along with the global rise in power of far-right populists.

such a hands-off approach is plainly in line with much of the Millian liberal theory. It is a position which Facebook seems to adopt only selectively; though the Community Standards suggest that the falseness of speech is not enough to justify limiting it, the company continues forward on the path of regulating misinformation, as seen in the case study included here. The Millian position that false speech as such does not warrant limitation is, then, not an extreme one, but rather one that is under-represented in the debate over speech regulation online.

Another significant challenge for online platforms in dealing with misinformation is the issue of distinguishing between opinion and fact. This question is crucial for the status of democratic politics and the right to freedom of expression, the latter being seen as a core civil right and a requirement of fair politics. My reflections on Arendt's work suggest that democratic politics depend on both the existence of factual truth and ability of citizens to voice their opinions, which may very well be related to false factual assumptions. If the political function of freedom of expression is to preserve social plurality and to balance the power of the majority, then minority opinions require careful protection. The danger of regulating misinformation, then, is the possibility that mere opinions or interpretations of fact are taken to be claims of fact and are therefore censored as "false" or "untrue". The accidental censoring of opinions has the potential to suppress politics; as seen in the case at hand regarding COVID-19 misinformation in France, the post in question included a critique of the government. In the liberal tradition, such critique of authority is not only tolerated but even encouraged in the pursuit of truth. Furthermore, due to the volume of online speech, it is likely that misinformation regarding topics of high public interest will be primarily subject to regulation; these topics, however, due to their public salience, are also likely to be part of political conversations where opinions and interpretations of facts are particularly important. Another issue which appears here is that, in regulating misinformation, social media platforms hold the power to choose which types of misinformation they choose to interfere with and which they choose to leave alone. Certainly the whole realm of human affairs is too broad to regulate in its entirety on the basis of fact. Arendt observes that factual truth has historically been relegated to the private sphere; if this is true, then it casts a doubt over the sustainability of efforts to require truthfulness in the public sphere today, which is more expansive and conflictual than ever.

The final issue which I believe to complicate the future of liberal freedom of expression online concerns the "less intrusive means" that Facebook (and other social media companies, like Twitter) have chosen to limit misinformation. My primary focus is on the approach of providing additional context and information along with posts that mention a topic of public interest. In the case of COVID-19, the Oversight Board mentions that, rather than remove posts thought to contain misinformation, Facebook's Community Standards provide the option of including additional context and information, such as is found in Facebook's COVID-19 Information Center. Before I comment any further, I would like to clarify that I do not necessarily mean to take a position against this means of responding to misinformation, but rather to explain and highlight its potential implications for liberal freedom of expression; this example, too, suggests that a liberal framework is not fitting for online speech.

The idea of this approach is, in cases of misinformation which do not reach the threshold of *imminent harm*, to provide links to authoritative sources of information. These sources may include national authorities, global organizations, or information pages created by Facebook (in the case of the COVID-19 Information Center). If these institutions and sources are not already considered authorities, they become concretely authoritative the moment that they are presented as the benchmark for factual comparison. This means of responding to misinformation is so unique to the contemporary world of information that it has not yet been thoroughly evaluated against the standards of liberal theory. Since it does not include the total removal of expression from a platform, it does not immediately signal an illiberal approach. However, from the perspective of doctrinal theory, it can be argued that this approach does not conform to liberal theory, especially concerning its assertion of objective "truth" in the public sphere and the resulting effects on the dynamics of authority between the individual and the government (or speech regulator, as in Facebook) regarding information.

T.M. Scanlon's theory of freedom of expression, drawn from foundational liberal theory, is relevant in the critique of this approach to misinformation. He warns against allowing a regulatory authority "the right to decide that certain views [are] false."[61] Here, it is important that Scanlon talks about false *views*, rather than false *claims of fact*. This is also central to my point. With Facebook's additional context and information approach, posts containing misinformation on certain topics of public interest will be accompanied by a statement from Facebook about the disputed nature of the information and a link to information from an authoritative resource. From a traditional liberal perspective, this approach is problematic when false claims of fact are presented as a part of, or alongside, an opinion (as they often, if not always, are). The inclusion of an "additional information" link and notification, while directed

[61] Scanlon, "A Theory of Freedom of Expression," 217. Also see the section "Scanlon's theory of freedom of expression" of the present work.

at a post's false claim of fact, is attached to the whole post, which may include not just claims of fact, but also opinions. While this approach does not explicitly say that certain views are false, it does imply it, or at least cast a doubt on the integrity of the view. In practice, social media platforms act as part of the public sphere where people go to hear opinions and develop their own. When opinions are constantly marked by "additional context" notifications which direct readers to established authorities, the potential for deliberation is limited because of the emphasis given to the voice of the authority.

It is possible that we acknowledge that the additional context and information approach to misinformation, as well as the other approaches mentioned here, are out of line with the liberal doctrine of freedom of expression yet accept them within an otherwise liberal system because they may be necessary to achieve other aims (in the case of COVID-19 misinformation, the aim of protecting public health). This involves the acceptance of some degree of paternalism, which is typically intolerable by liberal standards, in the governance of online speech. This seems to be, so far, what is playing out before us.[62] The introduction of this approach in the context of misinformation regarding certain topics of public interest may increase the public's tolerance to such measures; this is potentially problematic in future cases, where large companies are again responsible for deciding which issues are of sufficient public interest to apply similar regulatory measures. There is reason to believe that a society cannot, or should not, be liberal in some aspects and illiberal in others, as such a configuration would offend the basic political principles of liberalism. Below, in conclusion, I will suggest an alternative approach to regulating information and content on social media which is more in line with liberal theory, and which may serve as the subject for future theoretical research and proposals.

Political theory allows us to understand the assumptions about individuals, society, and power that lie underneath our systems of politics and governance, Here, evaluating the applicability of traditional liberal theory to the issue of regulating misinformation online helps us determine important considerations for the path forward for digital expression. I have aimed to show that digital media has transformed the functioning of the public sphere of speech, expression, and information-sharing such that it is no longer appropriate to carry-over all previously applied standards of liberal theory. To apply a Millian theory of expression to

[62] Social media companies utilized a similar approach to misinformation during the 2020 Presidential Election in the U.S. Here, the aim was to protect the integrity and functioning of the civic processes of democracy.

digital misinformation would require a laissez-faire regulatory approach which is, as of yet, an unpopular notion. The direction of misinformation regulation to this point reveals that the individualistic concerns of traditional liberalism may be giving way in the digital era, which may be seen to demand more collectivist concerns in order to prevent harm.

It is also possible that future regulation of speech online strives towards liberal ideals while applying non-traditional or new standards that respond directly to the challenges posed by digital media. Social media as the contemporary "public sphere" involves plenty of factors that influence autonomy and individual liberty which I have not yet had the space to discuss here. One area which I suggest as crucial to the future of a liberal culture online, and as an important topic for future research on the topic of digital expression, is that of the personal data, targeted ads, and content algorithms on social media. Social media companies—and the companies selling products on them—use personal data to create an online experience which is tailored to the individual. The idea is to show people things that they like and agree with; this benefits the social media company by ensuring that people enjoy and will continue using their service, and benefits companies advertising their products by finding the specific people more likely to purchase from them. Importantly, advertisement should not be separated from opinion or "speech" content, as political advertisements are part of the picture, and commercial actors inherently carry ideological assumptions which are communicated in their advertising. These normalized practices of promoting certain content and presenting tailored advertisements are in themselves a manipulation of the freedom of expression in that the process of opinion formation is automatically skewed.

If, from a Millian perspective, citizens have "the right to make up their own minds", then it can be argued that such a right depends more basically on the ability of citizens to exist within a digital public which is representative of the world and its plurality, rather than a curated, distilled version which is designed to appeal to and falls in line with an individual's existing opinions and preferences. Indeed, misinformation may be more likely to become integrated into the individual's worldview when it is read among similar content. When the body of content that individuals interact with on a daily basis is partially determined by their own preferences and behavior rather than only by their co-existence in a common world with other speakers, there is less opportunity for interaction with opposing opinions and ways of thinking. A traditional liberal (and Millian) defense of freedom of expression is that, for speech we disagree with, there is the opportunity of counter-speech, which strengthens the public discourse and requires that both sides truly engage with argument. With the current configuration of the digital public sphere, there is less room for counter-speech than in the pre- or non- digital sphere, where content may still be somewhat “curated”, but to a lesser extent. Offline, we can choose to read only certain publications and go to only certain public events, but social media quite literally provides us with a feed of hyper-curated content (some of which we choose and some of which is “suggested” to us) that exists largely in an abbreviated but more numerous form than information offline. In other words, the digital public sphere has become an environment which sacrifices the plurality of expression for economic interest and justifies this practice by claiming that it creates a better experience for social media users. Here, the connection to misinformation is subtle (if not indirect), but crucial—false claims of fact are bound to be more impactful on opinion and behavior in an environment where we are continually losing the opportunity and incentive to challenge ideas and hold a critical attitude to claims of fact. A digital public sphere structured by traditional liberal doctrine would involve greater restraint of social media companies and advertisers than it would regulation on the speech of individuals. The dominance of commercial interests online has created a digital public sphere that lacks the plurality and dynamics of communication of the non-digital world; this, in addition to the nature of digital speech, is to the effect that liberal standards such as the harm principle are no longer applicable, as the other characteristics of a liberal society are missing. Rather, a liberal approach to online speech requires new principles and approaches which look towards fostering plurality and limiting the power of social media companies not only in regulating content, but also in controlling the flow of information in a digital society.

Bibliography

Allan, Richard. "Hard Questions: Where Do We Draw the Line on Free Expression?" *Facebook, Inc.,* August 9, 2018.

Arendt, Hannah. "Truth and Politics." *The New Yorker*, 25 February, 1967.

Case Decision 2020-006-FB-FBR. Facebook Oversight Board (2020).

Judith Butler, Excitable Speech: A Politics of the Performative. New York & London: Routledge, 1997.

"Facebook Community Standards." *Facebook, Inc.*, 2021. https://www.facebook.com/communitystandards/

Feinberg, Joel. "Profound Offense." in *Mill's On Liberty: Critical Essays,* edited by Gerald Dworkin, 137-166. Maryland: Rowman & Littlefield Publishers, Inc., 1997.

Fukuyama, Francis. "Liberalism and its discontents: The challenges from the left and the right." *American Purpose,* 5 Oct. 2020. https://www.americanpurpose.com/articles/liberalism-and-its-discontent/

The United Nations General Assembly. 1966. "International Covenant on Civil and Political Rights." *Treaty Series* 999 (December): 171.

Knowles, Dudley R. "A Reformulation of the Harm Principle." *Political Theory* 6, no. 2 (May 1978): 233-246.

Langvardt, Kyle. "Regulating Online Content Moderation." *The Georgetown Law Journal* 106 (2018): 1353-1388.

Lyons, David. "Liberty and Harm to Others." in *Mill's On Liberty: Critical Essays,* edited by Gerald Dworkin, 115-136. Maryland: Rowman & Littlefield Publishers, Inc., 1997.

Mill, John Stuart. "On Liberty." In *On Liberty, Utilitarianism and Other Essays,* 5-112. Edited by Mark Philip and Frederick Rosen. New York: Oxford University Press, 2015.

Mulnix, M.J. "Harm, Rights, and Liberty: Towards a Non-Normative Reading of Mill's Liberty Principle." *Journal of Moral Philosophy* 6 (2009): 196-217.

"'Post-truth' declared word of the year by Oxford Dictionaries." *BBC News*, November 16, 2016.

Ramírez, Imanol. "Online Content Regulation and Competition Policy." *Harvard Law School Antitrust* Association, December 3, 2020.

Scanlon, Thomas. "A Theory of Freedom of Expression." *Philosophy & Public Affairs* 1, No. 2 (Winter 1972): 204-226.

Turner, Piers Norris. "'Harm' and Mill's Harm Principle." *Ethics* 124 (January 2014): 299-326.