UNIVERSITY OF RIJEKA
FACULTY OF HUMANITIES AND SOCIAL SCIENCES

Julija Perhat

THE CASE OF GENDER PEJORATIVES: THE SEMANTIC, ETHICAL AND POLITICAL DIMENSION

DOCTORAL THESIS

Rijeka, 2024

UNIVERSITY OF RIJEKA
FACULTY OF HUMANITIES AND SOCIAL SCIENCES

Julija Perhat

THE CASE OF GENDER PEJORATIVES: THE SEMANTIC, ETHICAL AND POLITICAL DIMENSION

DOCTORAL THESIS

Advisor: dr. sc. Elvio Baccarini

Rijeka, 2024

SVEUČILIŠTE U RIJECI

FILOZOFSKI FAKULTET

Julija Perhat

RODNE POGRDNICE: SEMANTIČKA, ETIČKA I POLITIČKA DIMENZIJA

DOKTORSKI RAD

Mentor: dr. sc. Elvio Baccarini

Rijeka, 2024.

Mentor doktorskog rada: dr. sc. Elvio Baccarini

Doktorski rad obranjen je dana _____ u/na

_____, pred povjerenstvom u sastavu:

1. _____
2. _____
3. _____
4. _____
5. _____

ACKNOWLEDGMENTS

First, I would like to start this section by a short *in memoriam* to my dear mentor who has been with me from the beginning of my academic career and who has mentored my BA and my MA thesis – the late professor Nenad Miščević. He was my mentor since the beginning of my PhD career and since the beginning of writing this thesis. Unfortunately, he passed away just months prior to submitting this thesis and, even though his name is not on the front page, I am indebted to him in every way. His loss is deeply felt, and I thank him for his guidance and friendship throughout the years.

I was lucky enough to have my dear co-mentor, and now mentor (even though I always considered him as such), Professor Elvio Baccarini, by my side from the beginning of writing this thesis who has advised me and guided me throughout all of this and for which I will be forever grateful.

I am indebted to all the friends and colleagues who have generously provided their support and insight: Enes Kulenović, Ivan Cerovac, Dan Zeman, Mirela Fuš-Holmedal, Andrej Jandrić, Martina Blečić, Sanja Barić and Snježana Prijić-Samaržija. Without their invaluable input and support this thesis would not be completed. I am also thankful to Monika Zeba for her time and effort poured into proofreading the thesis. I also want to thank all the colleagues who have listened to and commented on parts of the dissertation at conferences and workshops, especially to Miranda Fricker for her encouragement. Furthermore, I extend my gratitude to all my work colleagues, especially my friend and principal, Ana Tomić-Njegovan, for being immensely understanding, as well as to Milan Martuslović for helping with technical issues.

Finally, I thank my family; my husband and son, parents and all my (now late) grandparents, who have supported me throughout the years. I dedicate this to my son, Neo, and I hope he will be proud of me one day.

SUMMARY

By combining philosophy of language, political philosophy, and epistemology, the dissertation explores the ethical and political dimensions of slurs, areas that have been understudied despite recent increased philosophical interest. Slurs are a part of the pejorative cluster, and they express derogatory attitudes towards their targets. Slurs are the most used vehicle of hate speech which makes them a critical focus for understanding the broader phenomenon of hate speech. The main goal of the thesis is to address the harm caused by slurs that do not traditionally fit into the category of hate speech. Namely, some slurs do not legally qualify as hate speech but still cause significant harm, both individually and collectively. The most typical examples of these kinds of slurs are gendered slurs for women, such as *whore, slut*, etc. The idea that gendered slurs target women as a group and not just individuals has been challenged, and the thesis presents a possible answer to this challenge by augmenting existing theories of slurs. Namely, I utilize Miščević's (2016) idea of slurs having layers and suggest a new layer: the negative identity-prejudicial stereotype layer, inspired by Fricker's (2007) work. Finally, the pivotal aspect of the thesis is the introduction of the novel concept of derogatory-labeling injustice which explains how such slurs (prime examples being gendered slurs for women) cause significant harm through negative identity prejudice evoked by their literal uses. The dissertation, as stated, augments existing theories of slurs and highlights the role of negative identity prejudice in generating this kind of injustice. By examining slurs through the lens of philosophy of language, social and political philosophy, and epistemology, the thesis contributes to a deeper understanding of slurs and their impact.

Key words: slurs, hate speech, harm, identity prejudice, derogatory-labeling injustice

SAŽETAK

Kombinirajući filozofiju jezika, političku filozofiju i epistemologiju, disertacija istražuje etičke i političke dimenzije pogrda. Navedena područja nedovoljno su istražena unatoč nedavnom povećanom zanimanju za njih u filozofiji jezika. Pogrde podpadaju pod pogrdnice, a izražavaju negativne stavove prema žrtvama. Pogrde su najčešće korišteno sredstvo govora mržnje što ih čini kritičnom točkom fokusa za razumijevanje šireg fenomena govora mržnje. Glavni cilj ovoga rada jest istražiti štetu koju čine one pogrde koje se tradicionalno ne uklapaju u kategoriju govora mržnje. Naime, neke pogrde se pravno ne kvalificiraju kao govor mržnje, ali ipak uzrokuju značajnu štetu, sličnu onoj koju uzrokuje govor mržnje. Najtipičniji primjeri ove vrste pogrda su rodne pogrde za žene, kao što su *kurva*, *drolja*, itd. Ideja da su rodne pogrde usmjerene na žene kao skupinu, a ne samo pojedince, je dovedena u pitanje, a rad predstavlja mogući odgovor na taj izazov proširenjem postojećih teorija o pogrdama. Naime, koristim se Miščevićevom (2016) idejom da pogrde imaju slojeve te predlažem novi sloj, a to je negativni identitetsko-predrasudni stereotip, inspiriran radom Mirande Fricker (2007). Ključni aspekt disertacije je uvođenje novog koncepta nepravde – nepravde pogrdnog etiketiranja – koji objašnjava mehanizme takvih pogrda (glavni primjeri takvih pogrda su pogrde za žene). Odnosno, dokazuje se da takve vrste pogrda, u njihovoj doslovnoj upotrebi, uzrokuju značajnu štetu evocirajući negativne predrasude o identitetu. Disertacija, kao što je navedeno, proširuje postojeće teorije o pogrdama i naglašava ulogu predrasuda o negativnom identitetu u stvaranju nove vrste nepravde. Proučavajući pogrde kroz prizmu filozofije jezika, socijalne i političke filozofije te epistemologije, rad pridonosi dubljem razumijevanju pogrda i njihovog utjecaja.

Ključne riječi: pogrde, govor mržnje, šteta, negativne predrasude identiteta, nepravda pogrdnog etiketiranja

Contents

INTRODUCTION

"There is a word
Which bears a sword
Can pierce an armed man -"
Emily Dickinson *There is a Word*

This quote was also the beginning of my MA thesis written in 2012. The fact that I am inclined on using it again shows how little has changed in terms of my intuitions on how powerful words can be. As I explained in my MA thesis (2012), Emily Dickinson thought words to be so powerful that she found it fitting to compare them to a cold weapon, and the injuries they leave she compared to physical ones. It is probably true that at some point in our lives we have found ourselves very hurt by words, so perhaps we can all relate to Dickinson's feelings. When it comes to hurtful words, the most known association that comes to mind immediately is probably hate speech. Hate speech has been and remains at the center of attention in many fields of study, from philosophy to law and politics, and for a good reason. The reason why hate speech remains to be a hot topic across various fields of study is because language can be a powerful tool. We need only to remind ourselves how using language for certain agendas can be deadly. For instance, during the Middle Ages accusing somebody of being a witch could have potentially led to death by burning. There are numerous examples where language played a crucial role and where words presented danger. Hate speech may take many forms—it can be expressed through acts (such as cross-burning which represents racial hatred) or signs (such as the swastika, the symbol of the Nazis). However, one of the most used ways to express hatred is through words and derogatory words will be the focus in this thesis.

Although, as emphasized, hate speech can take many forms, in this thesis I will focus on one particular aspect of derogatory language that has puzzled philosophers of language for considerable time—slurs. So, what are slurs? Slurs are pejoratives, or, as Christopher Hom (2010) put it, they are a cluster of pejoratives and pejoratives "express the derogatory attitudes of their speakers" (Hom 2010, 164). Jeshion nicely summarizers the purpose of slurs: "to signal that their targets are unworthy of equal standing or full respect as persons, that they are inferior as persons" (Jeshion 2013a, 232). Slurs are a complex phenomenon and philosophers of language have tried to explain their pragmatic and semantic elements in order to fully understand how they work. However, the research hasn't yielded conclusive results. While researching the content of slurs, one area of research has been neglected, and that is the ethical and political effects slurs generate. Although, as of recently, many more

philosophers have included the ethical and political dimension in their study of slurs, the area still remains understudied. My work in this thesis will focus precisely on these dimensions—the ethical and the political ones. Thus, the main goal of this thesis is to contribute to the debate on hate speech and to explore the understudied ethical and political dimensions of the most used vehicles of hate speech—slurs. I will delve into the complex interplay between slurs, hate speech, prejudice and potential harm these may cause, thus combining and providing an outlook to these issues from various disciplines, namely philosophy of language, social and political philosophy and epistemology. To fully understand hate speech and slurs, I think it is necessary to explore and combine these fields in order to capture the full scope of their effect.

The natural question to perhaps ask at this point is why focus specifically on slurs when hate speech may have a much broader scope (hate speech can be expressed through signs, acts and so on). The first part of the answer to this question is that people rely on language to communicate and communication is an essential part of being human. People rely on language for various reasons: to gain knowledge (Fricker 2007), to get their meaning across, to form connections with others, to be a part of a community, etc. Derogatory language such as slurs are a disturbance in communication and they are the most often used vehicle of hate speech. As such, it makes sense to turn our focus towards such phenomenon. The second part of the answer is that, on the one hand, by examining how slurs work, we can extrapolate these findings to understand other aspects of hate speech better. On the other hand, one of the main points I will make is that there are some slurs that do not fall into the category of hate speech and, as far as the legal domain is concerned, they wouldn't be considered hate speech. However, even though they wouldn't be considered hate speech in most cases, I claim they still harm their targets the same way hate speech does, both individually and collectively. Therefore, the arguments I present from other esteemed authors on the harm hate speech does will be used for slurs as well. It is perhaps useful at this point to consider the demarcation between slurs and hate speech.

As will become evident in Chapter I, there is no universal definition of hate speech, and the legal treatment of hate speech varies across countries. Nonetheless, by examining different definitions of hate speech, one can derive some common characteristics. Namely, most definitions of hate speech consider the target groups to be groups of people with ascribed characteristics and the focus is on public speech. These characteristics of hate speech will be accepted in this thesis as well. As mentioned, even though when thinking about hate speech slurs most often come to our mind, there are slurs that wouldn't be considered hate speech within most laws. But, even so, I claim these cause the same kind of harm hate speech does: they cause harm not only to the individual, but to the entire group as well. The prime examples of slurs that are usually not considered hate speech are gendered slurs for women,

such as *whore*, *bitch*, *slut,* etc. The demarcation between slurs and hate speech could be portrayed in the following manner:

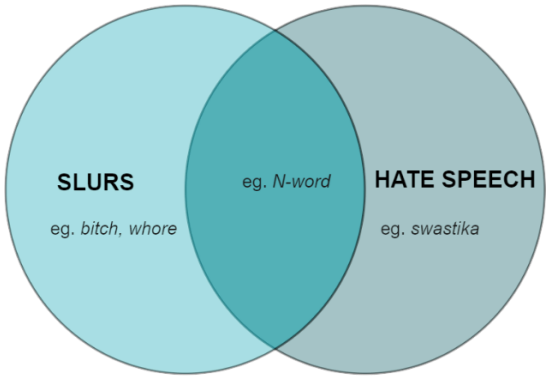

Figure 1

I understand hate speech as a broad concept encompassing acts (such as cross burning), or symbols (such as the swastika). The intersection of slurs and hate speech includes words such as the N-word which most laws traditionally recognize as hate speech. However, as noted above, there are slurs that wouldn't be considered hate speech (prime examples being gendered slurs for women). Nevertheless, even those slurs that wouldn't be recognized as hate speech by most courts, still cause the same harm ascribed to hate speech.

That said, throughout this thesis I will provide arguments by esteemed authors on harm caused by hate speech. I utilize these arguments and extrapolate them to a broader scope than what is considered hate speech in most cases. Namely, whenever I mention these arguments on hate speech, I will use them to argue that slurs, even when not considered hate speech, cause the same harms.

In fact, this brings us to the pivotal aspect of this thesis: the introduction of the novel concept of *derogatory-labeling injustice* (inspired by Fricker's and Kukla's notions). This kind of injustice hasn't been described in literature so far, but it was foreshadowed by Fricker (2007) when she said that the tracker prejudice she identified may give rise to various kinds of injustices. Even though some slurs will not fall into the category of hate speech, their systematic use will cause various harms not only to an individual, but also to a group the individual belongs to. This harm amounts to what I call derogatory-labeling injustice, namely an injustice that happens in a discourse setting where the speaker, who is in a position of power, by using derogatory language, i.e., slurs, labels the target with negative identity prejudice, and thus wrongs the target by harming one or more of their important interests. The main driving force behind derogatory-labeling injustice is Fricker's notion of negative identity prejudice. When a slur is uttered, these negative identity prejudices are evoked.

Slurs possessing and evoking stereotypes is not a new idea as it has been put forward by various authors, such as Williamson (2009), Hom (2008), Camp (2011), Miščević (2016), and others. I aim to offer, not a novel theory of slurs, but a modification of these views. I claim that slurs evoke a specific kind of prejudice—the identity prejudice described by Fricker. Or, more precisely, a negative identity-prejudicial stereotype defined as follows: "A widely held disparaging association between a social group and one or more attributes, where this association embodies a generalization that displays some (typically, epistemically culpable) resistance to counter-evidence owing to an ethically bad affective investment" (Fricker 2007, 35). The slurs that evoke these kinds of prejudices would be slurs used in their literal sense to degrade the target, something that Jeshion (2013a) refers to as weapon-uses of slurs. To explain how negative identity prejudices apply to slurs, I utilize Miščević's (2016) idea of slurs having layers. According to him, there would be five levels of slurs: causal-historical, minimal descriptive, negative descriptive evaluative, prescriptive, and expressive. I adapt his view and introduce the negative identity prejudice layer which would, in my view, ground the normative evaluative judgment of the target. Even though I adapt Miščević's idea of layers, the issue of demarcation between layers, i.e., whether each layer would be a part of pragmatics or semantics of a slur, I leave open for some further research. This particular problem, even though rightly important in philosophy of language, is not crucial for the main aspect of my claim, which is that the systematic use of slurs gives rise to derogatory-labeling injustice.

By combining philosophy of language, political philosophy and epistemology, the dissertation offers a new perspective on already fairly researched phenomena of hate speech and slurs. Since there are slurs that traditionally don't fit into the category of hate speech, but it still seems they harm their targets, the dissertation provides an answer to this question. By introducing the novel concept of derogatory-labeling injustice, it becomes clear how slurs that are not hate speech may cause significant harm. The underlying root of this harm are the negative identity prejudice introduced by Fricker (2007) which are evoked by the literal uses of slurs. Thus, the dissertation, as well as offering a novel concept of injustice, offers an augmentation of existing theories of slurs and identifies negative identity prejudice as being one of the layers of slurs that generates this kind of injustice.

The dissertation is divided into five chapters.

The first chapter deals with questions from political philosophy, namely the debate on hate speech and freedom of speech. In order to delve deeper into the topic of slurs and harm, I first present an overview of the debate that has puzzled political philosophers for decades. In liberal societies one of the fundamental principles that we want to protect is freedom of speech. As such, freedom of speech is protected by the UN's Universal Declaration of Human Rights where Article 19 guarantees freedom of opinion and expression which includes "freedom to hold opinions without interference and to seek, receive and

impart information and ideas through any media and regardless of frontiers" (The UN Universal Declaration of Human Rights). World's democracies have also sought to protect freedom of speech. In the US it is protected by the First Amendment where it is stated that "Congress shall make no law…abridging the freedom of speech, or of the press", and in Croatia freedom of speech is also guaranteed by our Constitution. Since freedom of speech is a fundamental principle we want to protect, the question is whether we have the right to, and if so, in which cases, to restrict some forms of speech? Many authors argue that, in order to restrict speech, the speech needs to cause harm. In this Chapter, I will present the most prominent arguments from various authors that advocate for some kind of restriction of hate speech. Usually, authors that argue for some kind of restriction of hate speech, stress the possible harm hate speech incurs. I will agree with those authors, and I will utilize their arguments to claim that there are parts of speech that are usually not considered hate speech, but that also cause these harms to their targets. However, the authors who claim that hate speech causes harm have struggled to provide evidence of its effect on the targets. Some authors have called for the need of pointing to empirical evidence of the harm caused by derogatory language (Simpson 2019; Heinze 2016). I partly agree with this and I utilize empirical evidence on derogatory language to pinpoint the harm being caused to targets.

The second Chapter, therefore, is a background needed to move forward in the investigation. My main claim is that there are some parts of speech, i.e., slurs, that do not fall into the category of hate speech but that still cause significant harm to the target. This is due to slurs evoking what Fricker (2007) described as negative identity prejudice. Since I partly agree with the claim that the harm done by hate speech somehow needs to be backed up by empirical evidence, in the second Chapter I present empirical evidence on how stereotypes and prejudice may affect the targets of derogatory speech. I utilize these findings to show how evoking negative identity prejudice may negatively affect the targets of slurs. Establishing identity prejudice as being a part of the slur (semantically or pragmatically) provides us with a "hook" to trace back the harm that has been done by the systematic uses of slurs. Finally, in the last part of the second Chapter, I give a brief overview of Miranda Fricker's (2007) work on testimonial injustice since I rely on her work in order to better understand how slurs work.

With the needed background on how stereotypes and prejudice may affect us set, in the third Chapter I turn my focus to slurs. As already emphasized, there are some slurs that wouldn't fall into the category of hate speech. Despite that, they still cause significant harm to its targets. In this Chapter, I analyze how this is possible. In order to better understand what slurs are, in this Chapter I analyze some of the examples of slurs with the help of dictionary entries. Then, I turn my focus to prime examples of slurs that wouldn't traditionally fall into the category of hate speech, and these are gendered slurs for women, such as *whore*, *slut*, etc. In most cases, these slurs wouldn't be considered hate speech by

most courts, they would rather fall into the category of libel, for example.[1] Some authors argue that gendered slurs are not slurs at all since they do not target women as a group (Nunberg 2018), and some argue that they indeed are slurs, but that they are inherently different than racial slurs in that it seems they lack a neutral counterpart which is usually a part of the definition of slurs (Ashwell 2016). For answers to these questions, I turn to Justina Diaz Legaspe (2018) who has offered what seems to be a satisfactory explanation to points raised by Nunberg and Ashwell, and I will accept and build on her response. Introducing negative identity prejudice into the picture may help strengthen Legaspe's points. To explain how gendered slurs work, she introduces P-behavior, a behavior that is not deemed acceptable according to the established norm (Legaspe 2018). Legaspe (2018) claims that every member of a group labeled with a gendered slur can *potentially* exhibit P-behavior. When somebody refers to a woman as a *whore*, on the face of it, it may seem they are saying something only about the individual. However, understanding gender slurs for women requires a deeper understanding of our patriarchal society. Introducing identity prejudice explains the reason why all women have the potential to P-behave: because identity prejudice applies to all women. The punchline that strengthens Legaspe's (2018) view is this: *women should not P-behave because of identity prejudice*. There is an underlying identity prejudice that applies to all women—all women should aspire to be lady-like and should be ashamed of their sexuality. This is also the main augmentation I aim to make to existing theories of slurs in this chapter: slurs evoke negative identity prejudice. To explain how negative identity prejudice applies to slurs, I utilize Miščević's (2016) idea of layers. These layers would be a neural counterpart, a negative evaluative judgment of the target, a negative identity prejudice layer, a historical link, a feeling of contempt and the issue of epistemic culpability. Introducing negative identity prejudice as being part of the slur provides us with an explanatory advantage of a slur's content: the normative evaluative judgment is grounded in negative identity prejudice.

After establishing that slurs evoke negative identity prejudice when uttered, I turn to the fourth Chapter and the introduction of the pivotal aspect of this thesis, which is the novel kind of injustice: derogatory-labeling injustice. I first start by making a connection between Fricker's (2007) testimonial injustice and slurs and I utilize some of the notions mentioned by Fricker and apply them to slurs. Namely, I utilize Fricker's notion of social power in that I claim that each discourse has a certain power relation between the speaker, the listener, and the target. When the speaker is in a position of power, the perlocutionary potential the slur has is even greater. In fact, this means that slurs create an atmosphere that is fertile ground for testimonial injustice to take place. Secondly, I utilize Fricker's idea of epistemic culpability. The speaker is epistemically culpable when using slurs in their literal sense to degrade. Moreover, if the speaker is in the position of power, such as a political one, we can

[1] A thanks goes to Enes Kulenović for drawing my attention to this.

hold them even more accountable due to their social position and the social power they hold. Finally, I move on to the central case of this thesis—to examine what slurs do when uttered. I will claim that slurs enact social harms on the target and that these harms can be manifested in two effects: the primary and the secondary effect, where the primary effect is more immediate and the secondary effect is more consequentialist in nature in the sense that it may have a long-lasting effect. In order to reach a full understanding of what slurs do when uttered and what kind of harm they cause, it is best to review these effects from various perspectives: the speaker's, the listener's, and the target's. Until now, this particular method of considering all of the perspectives in a discourse has not been employed in investigating the impact of slurs. Furthermore, these harms can be traced back to the effect of prejudice discussed in Chapter II. The main claim is that even slurs that are not considered hate speech have a profound negative effect on their targets and on the society as a whole. I discuss several harms that have a long-lasting effect amounting to the secondary effect of the slurs. These are: a) stereotype threat, b) self-fulfilling prophecy, c) maintaining status quo, d) hindering deliberation, e) impeding opportunities to acquire primary goods, and f) hindering thinkers' interests. Here, I also examined some harms that haven't been discussed in literature before (as far as I know of), namely the ability to impede opportunities to acquire primary goods and hindering thinkers' interests (as an answer to Seana Shiffrin's thinker-based account). Finally, I am able to provide a definition of the novel concept of derogatory-labeling injustice: *derogatory-labeling injustice happens in a discourse setting where the speaker, who is in a position of power, by using derogatory language, i.e., slurs, labels the target with negative identity prejudice, and thus wrongs the target by harming one or more of their important interests*. This kind of injustice provides an explanation as to why it seems that even though some slurs are not traditionally considered hate speech, they cause harm to their targets. Considering this novel kind of injustice, we can now better understand the three concepts:

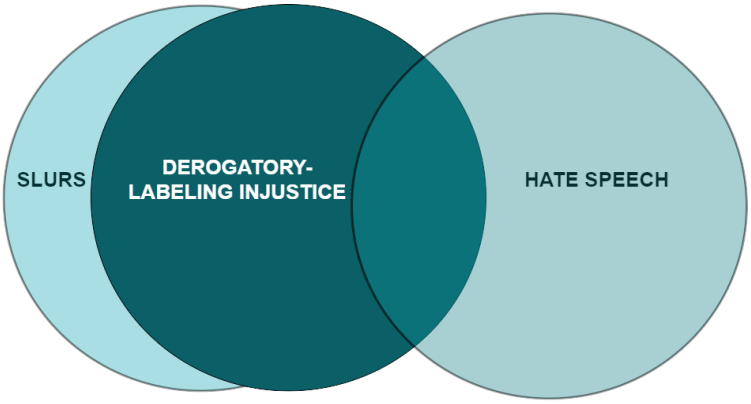

Figure 2

This shows us that, on the one hand, there is the concept of hate speech which I understand in broader terms, i.e., it can be expressed by symbols, or acts. However, slurs that are not considered hate speech also cause harm to their targets, or, more specifically, by evoking negative identity prejudice, they cause derogatory-labeling injustice which, in part, may overlap with the concept of hate speech. This means that, in some cases, slurs can be considered hate speech and also cause derogatory-labeling injustice. In other cases, slurs may not be considered hate speech, but, nonetheless, they may cause derogatory-labeling injustice. Finally, there will be slurs that are used in their literal sense to degrade the target, but they will not cause derogatory-labeling injustice (e.g., the Croatian word *Tovar* used to refer to football fans from the Croatian town of Split). This is because for derogatory-labeling injustice to take place certain conditions have to be met: it is produced by the literal and systematic use of slurs, where the targets are historically marginalized groups that are already in a disadvantaged position in society, and, finally, it has to be used by a person who bears a certain amount of social power.

Finally, in the fifth and last Chapter, I ask what can be done with such phenomenon in our language. In this last Chapter, I explore some possible responses to hate speech and derogatory-labeling injustice where I claim that the optimal solution could be found by coordinating several strategies. In other words, in order to be effective, the responses to such phenomena have to be two-fold, i.e., they have to come from two directions. Thus, I divide the possible responses into two areas: the first set of responses are the ones that can be given by members of society on an individual level and these will be concerned with epistemic responsibility, and the second set of responses are concerned with institutionalized responses where I argue that the state has the responsibility to react to harms done to its most vulnerable members by providing certain institutionalized protections. These two sets of responses complement each other, and I find this to be the most effective way to combat injustices produced by prejudice.

CHAPTER I: FREEDOM OF SPEECH AND HATE SPEECH

1.1. Introduction

In this Chapter, I aim to give an overview of a debate surrounding the notion of freedom of speech and hate speech. Freedom of speech and hate speech are two intertwined notions, and the debate about one cannot go without the other. Freedom of speech is a fundamental right that is protected in all liberal democracies as it is essential for other rights such as freedom of assembly, freedom of religion, and so on. But there are instances where it seems that absolute freedom of speech causes some harm and presents a danger. The question is then, do governments have the right to limit such an important right, and if so, in which cases? I will give a brief overview of the debate since it is important for the later in-depth discussion of hate speech and injustices that stem from it. After pointing out the most known arguments for the protection of free speech, I will consider arguments from the opposition, i.e., those that are in favor of restricting free speech. In fact, most rights concerning our freedom to do something are in some sense restricted. For example, freedom of movement is restricted by property ownership, state borders, or, as we have recently witnessed, extraordinary conditions such as a pandemic. The same rationale applies to freedom of speech. There are instances where restriction is needed. However, where the line is drawn depends on our interpretation of freedom of speech. This debate is very broad thus encompassing all views is an impossible task for the current project, so I will address just the most prominent views from each side of the debate. I will side with the authors who think hate speech should be regulated and sanctioned in some sense, and in later chapters, I will present a rationale for responding to such harmful speech. The emphasis here is precisely on the harmful part; I will claim that certain speech can cause serious harm that needs to be ameliorated. That is why it will be important to build a distinction between harm and offense in this chapter where I will claim that mere offense will not be enough to restrict any kind of speech. After presenting the debate about the freedom of speech and hate speech, as well as the offense versus harm debate, I will give a brief overview of the current affairs considering hate speech. Namely, I will provide some of the most known definitions of hate speech and try to pinpoint some joint characteristics. Since defining hate speech and the legal treatment of it are connected, I will juxtapose the legal treatment of hate speech in the US and Europe. This will show how the confusion surrounding the theoretical understanding of hate speech seeps into the legal system so that there is no unified legal treatment of hate speech. This shows that there is a need for better understanding of all the mechanisms underlying and surrounding hate speech. In the later Chapters, I will try to clarify some

of the issues surrounding hate speech, and I will do so by focusing on one important aspect of hate speech—slurs. Furthermore, I will extrapolate the arguments that claim hate speech causes certain social harms and claim that they can be applied to slurs, as well.

1.2. Debate on freedom of speech: a preview

As previously mentioned, freedom of speech has a profound value in liberal democracies. As such, it is protected by the UN's Universal Declaration of Human Rights where Article 19 guarantees freedom of opinion and expression which includes "freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers" (The UN Universal Declaration of Human Rights). Every liberal democracy has, in one way or the other, protected freedom of speech. The US, for example, has done so under the First Amendment which states that "Congress shall make no law…abridging the freedom of speech, or of the press", and in Croatia freedom of speech is also guaranteed by our Constitution. But, as Simpson (2013) points out, even though liberals want to guarantee and protect freedom of speech, they also "want to use the disciplinary function of the law to combat and reform identity-based social hierarchies", therefore "the twin liberal commitments to free speech and social equality thus seem to come into conflict where hate speech is at issue" (Simpson 2013, 2).

There are various arguments why freedom of speech is so important and why it should not be restricted, but the most known arguments fall into three camps: those which focus on gaining the truth, those which focus on autonomy[2] and those which focus on democracy. Recently, there has been what can be considered a new approach from a thinker's perspective proposed by Seana Shiffrin (2014) which she calls the thinker-based approach. I will briefly present the most prominent arguments from these camps and then contest them with the most prominent criticism. I will leave the criticism of the last thinker-based approach for the latter chapters since its criticism needs some preconditions, such as understanding the effect of hate speech, which I hope to accomplish in future chapters.

One of the probably most known advocates of freedom of speech is John Stuart Mill. In his essay *On Liberty,* he pointed out that it is important not to censor speech since it can lead us to truth. Mill summarizes his argument as follows: a) the opinion which is censored might be true, b) even if the censored opinion is false, it may be partially true, and c) if the opinion is the whole truth, if not contested it may lose its power and even become a dogma, i.e., mere formality we follow blindly (Mill 1879, 25). So, by censoring

[2] When it comes to arguments from autonomy, we can also divide them into two camps: those that focus on the listener's autonomy (Scanlon 1972, Dworkin 1996, Nagel 2002, etc.) and those that focus on the speaker's autonomy (Baker 1989, Cohen 1993) where the listener based approaches are focused on the ability to access information and the speaker based approaches are focused on being able to freely articulate one's ideas. Although there are compelling arguments on both sides, due to the limitations of this thesis, I will present the most prominent one proposed by Scanlon.

speech we risk losing knowledge. He supports his argument by reminding us that, in the past, there have been numerous occasions where people held false beliefs and deemed the actual true ones false and crazy. Furthermore, Mill claims that it is of the utmost importance to allow discussion of various opinions because the very discussion gives power to our opinion if it is not refuted. By allowing the opinions we hold to be true to be contested, we are justifying our beliefs if it turns out they are not proven wrong. For us to grow as persons and to become wiser, we need to allow our opinions to be contested and, if needed, to be set right. But even Mill sets boundaries as to what one may say in certain situations. This is also one of the most known instances of when we have the right to restrict one's actions and it is formulated in his harm principle, which states that "the only purpose for which power can be rightfully exercised over any member of a civilized community, against his will, is to prevent harm to others" (Mill 1879, 10). In the third chapter of *On Liberty*, Mill provides us with examples of when this principle ought to be applied to restrict speech. According to Mill "even opinions lose their immunity, when the circumstances in which they are expressed are such as to constitute their expression a positive instigation to some mischievous act" (Mill 1879, 39). He supports this with an example of stating that corn dealers are starvers of the poor which should be allowed to circulate in the press as a free opinion, but that would not be the case if the same statement were to be uttered in front of an angry mob waiting outside the house of a corn dealer (Mill 1879). So, "acts, of whatever kind, which without justifiable cause do harm to others may be, and in the more important cases absolutely require to be, controlled by the unfavorable sentiments, and, when needful, by the active interference of mankind. The liberty of the individual must be thus far limited; he must not make himself a nuisance to other people" (Mill 1879, 39). As it is evident from the text, Mill was also of the opinion that in some instances, namely, when it presents imminent danger and harm to others, speech may be restricted. But, he is also clear in that mere offense is not enough to restrict speech since the offense is similar to simply having different tastes and one cannot be punished for that. There have been numerous interpretations of his work and the harm principle, but one that I am inclined to support is presented by David O. Brink in his article *Millian principles, freedom of expression, and hate speech*. Brink states that what is important for Mill is the ability to develop our deliberative capacities which would then manifest in a good human life (Brink 2001). In order to develop our deliberative capacities, we need to be able to have access to various liberties, such as liberty of thought and action. Brink builds upon this notion of importance of these capacities and claims that "because it is the importance of exercising one's deliberative capacities that explains the importance of certain liberties, the usual reason for recognizing liberties provides an argument against extending liberties to do things that will permanently undermine one's future exercise of those same capacities" (Brink 200, 138) and asks "can hate speech regulation be shown to protect or advance the very deliberative values that explain why

censorship is usually impermissible" (Brink 2001, 138)? It is Brink's understanding that since deliberative capacities are so important in Mill's account, some form of regulation of hate speech would be permissible. In other words, "if hate speech retards deliberative values, and hate speech regulation protects deliberative values, then we should not see hate speech regulations like the Stanford and neo-Stanford provisions as restricting fundamental liberties. Hate speech regulation can be seen as a well-motivated exception to the usual prohibition on content-specific regulation of speech" (Brink 2001, 142). As Brink explains, hate speech has a negative effect on our deliberative interests. The speaker disrespects her target but provides no explanation of her attitudes, leaving little room for debate. Also, emphasizes Brink, the usual immediate response to hate speech is violence or silencing. To further explain this, he turns to Charles Lawrence's (1993) explanation of visceral and inarticulate responses generated by hate speech, where Lawrence states how hate speech invokes a psychological reaction in a victim that disables the victim from any reasonable response. Thus, concludes Brink, "insofar as hate speech, like fighting words, expresses visceral attitudes and elicits inarticulate reactions, it doesn't engage deliberative values central to Millian and constitutional principles that normally protect speech" (Brink 2001, 140). Moreover, in order to engage in meaningful deliberation, certain conditions of mutual respect have to be met. But, as Brink claims, some empirical evidence[3] suggests that hate speech contributes to the feeling of disrespect and inadequacy in the targets thus discouraging them from participating in deliberation. Brink concludes that "because of the importance of deliberative interests in Mill's account of human happiness and in specifying fundamental interests and liberties, the adverse effects of hate speech on the deliberative interests of targets ought to be reckoned as harms" (Brink 2001, 146).

A different approach in defending free speech and any restrictions on it is built on autonomy. Even though there are various conceptions of autonomy, it can, in the broadest sense, be defined as the ability to make one's own decisions according to one's own will. One of the most prominent defenders of freedom of speech is Thomas Scanlon, who built his principle for defending free speech on Mill, therefore calling it the Millian principle that states:

> 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

[3] In the later Chapters, I will devote more attention to the empirical evidence since it is important for our task of trying to show how hate speech can affect us.

> 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. (Scanlon 2003, 14)

He defends this principle based on his account of autonomy, which he defines as being able to decide what to believe and how to act in the sense that a person needs to be able to decide for herself whether she is going to perform a certain action that is required by law. This principle is quite speech protective, especially combined with the view Scanlon pulls from Kant about the idea that the legitimacy of a government is dependent on the ability of citizens to regard themselves as equal, autonomous, rational agents (Mackenzie and Meyerson 2021). In the case of hate speech, the one responsible should be the listener if she is persuaded by harmful speech and decides to act on it. Scanlon later backtracked on his previous justification of the Millian principle because he recognized that it produces implausibility such that, according to the principle, false advertising would be permissible. Thus, Scanlon rejected the idea that autonomy puts an absolute constraint on the state's exercise of power (Mackenzie and Meyerson 2021) but "Scanlon continues to defend a theory of free speech which is autonomy-based, in the sense that he believes that the reason to protect freedom of speech is that it protects and advances important interests in substantive autonomy—not only of audiences in deciding what to believe and what reasons to act on, but also of speakers who seek to communicate and express their values to others, and bystanders or members of the general public who benefit from living in a society that enjoys freedom of speech" (Mackenzie and Meyerson 2021, 65). But, in his later views, he also admits that perhaps being shielded from certain bad influences (such as false advertising) can even better our autonomy.[4]

Arguments from democracy stress that in a democratic society citizens should be free to debate anything they see fit. If we want to continue to live in a democracy, we have to embrace viewpoint neutrality, otherwise the very foundation of democracy would be at risk (Heinze 2016). For democracy to function properly, the voters must be adequately informed on various matters, and that can be accomplished only by securing freedom of expression. Or, as Alexander Meiklejohn puts it: „When a free man is voting, it is not enough that the truth is known by someone else, by some scholar or administrator or legislator. The voters must have it, all of them... That is why no idea, no opinion, no doubt, no belief, no counter belief, nor relevant information, may be kept from them" (Meiklejohn 1948, 88). In addition to being an important means for people to be informed, freedom of speech also has an important function as a means for promoting

[4] See also Fish (1994).

political legitimacy, as Bhagwat and Weinstein (2021) emphasize. When citizens have an opportunity to be a part of a decision-making process where they can express their views, they tend to be more ready to obey the law (Bhagwat and Weinstein 2021).

However, as Bhagwat and Weinstein (2021) and Jeffrey W. Howard (2019) stress, the view on the restriction of speech will depend on the theory of democracy we endorse. For example, as Bhagwat and Weinstein (2021) emphasize, deliberative democracy will need a very broad protection of freedom of speech.[5] Or, as Howard (2019) points out, instrumental democrats might welcome restrictions on speech if it proves to be harmful to just outcomes. He adds that committed democrats need not think that every democratic decision has value, so perhaps one can conclude that restricting hate speech is a decision that would not reduce anything valuable (Howard 2019). Or, as Bhagwat and Weinstein (2021) conclude, "while free expression is necessary for democracy, the actual protection of free expression required by democracy is limited" (105).

Andrew Reid (2020) in fact makes a compelling argument by contesting the legitimacy argument, as he calls it, made by Dworkin (2009) and Weinstein (2017), which states that democratic legitimacy may be challenged by regulating hate speech. Reid (2020) indeed agrees that it may be the case that regulation of hate speech in a way harms democratic legitimacy, but he also argues that hate speech can have a negative effect on legitimacy, so there are strong reasons to support both claims. So, the question is—how can hate speech thwart political discourse? Reid provides two possible reasons:

> On an individual level, the targets of hate speech might be less inclined to participate in politics because of a sense that they are not treated with dignity, or in extreme cases because they feel threatened (Brown 2017, pp. 609–610). In this article, I focus, instead, on the claim that hate speech might have a deeper pathological effect on political discourse, because it might plausibly cause people to be taken less seriously in politics when they do decide to participate. The account of legitimisation that I have drawn upon in this article has been one that depends on participants in politics adhering to certain basic deliberative norms, specifically treating others with respect as deliberators. This mutual respect might be undermined if certain groups are routinely treated as if they lack core deliberative capacities, which is a potential effect of hate speech. (Reid 2020, 187)

For Reid, it is necessary to understand that the obligation to respect others in deliberation is difficult to achieve and should therefore be understood in terms of an ideal

[5] First developed in Dworkin (1996).

theory, meaning that the ideal cannot be met but that we should work towards it, for example, by enforcing norms in certain contexts. As Reid explains:

> Unlike the effects of censorship, which are direct, the harmful effects of hate speech on democracy only occur in a given context where there are existing inequalities and injustices. Hate speech does not undermine deliberative norms by its nature or as an intrinsic by-product of its content because such norms are independent of the content of the views expressed in discussion. Instead, hate speech can contribute to a moral environment where some citizens lose an effective political voice as a result of the way that democratic norms are reshaped. If the other conditions assumed in the ideal theory of legitimacy persisted, the state would be able to mitigate or compensate for these effects. This is not the case under nonideal conditions where social stigmatisation affects standing in politics in a more profound way. (Reid 2020, 192)

According to Reid, restrictions might be justified in cases where there are no formal obstacles for citizens to participate in deliberation, but where hate speech causes them to "participate less effectively as a result of changes in others' behavior and political norms" (Reid 2020, 188). But, still, there remains a worry that any restrictions on hate speech will harm the political process, so Reid (2020) cautions that "the legitimacy of state interference to limit hate speech must therefore be judged on a case-by-case basis" (189). Nonetheless, Reid (2020) concludes that we should try to work towards the democratic ideal and thus „in the case of hate speech, we might have reason to restrict freedom of expression because permitting it causes greater harms to others, or damages the process of the justification of laws more, than restricting them would" (195).

The last argument for freedom of speech we will look at is a thinker-based argument proposed by Seana Shiffrin (2014).[6] According to her approach, freedom of speech is crucial in developing ourselves as thinkers. Shiffrin views us as distinctive individuals with certain moral, rational, emotional, perceptual, and sentience capacities that have certain interests. Shiffrin identifies these interests as being:

> a. *A developed capacity for practical and theoretical thought*. Each thinker has a fundamental interest in developing her mental capacities to be receptive of, appreciative of, and responsive to reasons and facts in practical and theoretical thought, i.e., to be aware of and appropriately responsive to the true, the false, and the unknown.
>
> b. *Apprehending the truth*. Each thinker has a fundamental interest in believing and understanding true things about herself, including the

[6] Shiffrin's arguments are inspired by Mill (1879).

contents of her mind, and the features and forces of the environment from which she emerges and in which she interacts.

c. *Exercising the imagination*. In addition, each thinker has a fundamental interest in understanding and intellectually exploring non existent possible and impossible environments. Such mental activities allow agents the ability to conceive of the future and what could be as well as what could have been. Further, the ability to explore the nonexistent and impossible provides an opportunity for the exercise of the philosophical capacities and the other parts of the imagination.

d. *Moral agency*. Each thinker has a fundamental interest in acquiring the relevant knowledge base and character traits as well as forming the relevant thoughts and intentions to comply with the requirements of morality. (This interest, of course, may already be contained in the previously articulated interests in developing the capacity for practical and theoretical thought, apprehending the truth, and exercising the imagination [a–c].)

e. *Becoming a distinctive individual*. Each thinker has a fundamental interest in developing a personality and engaging more broadly in a mental life that, while responsive to reasons and facts, is distinguished from others' personalities by individuating features, emotions, reactions, traits, thoughts, and experiences that contribute to a distinctive perspective that embodies and represents each individual's separateness as a person.

f. *Responding authentically*. Each thinker has a fundamental interest in pursuing (a–e) through processes that represent free and authentic forms of internal creation and recognition. By this, I mean roughly that agents have an interest in forming thoughts, beliefs, practical judgments, intentions, and other mental contents on the basis of reasons, perceptions, and reactions through processes that, in the main and over the long term, are independent of distortive influences. In saying these processes are independent of distortive influences, I mean that the choices of what to think about and the contents of one's thoughts do not follow a trajectory fully or largely scripted by forces external to the person that are distinct from the reasons and other features of the world to which she is responding *qua* thinker. So, too, thinkers have an interest in revealing, sharing, and considering these mental contents largely at their discretion, at the time at which those contents seem to them correct, apt, or representative of themselves, as well to those to whom (and at that time)

such revelations and the relationship they forge seem appropriate or desirable. These are the intellectual aspects of being an autonomous agent.

g. *Living among others*. Each thinker has a fundamental interest in living among other social, autonomous agents who have the opportunities to develop their capacities in like ways. Satisfaction of this interest does not merely serve natural desires for companionship but also crucially enables other interests *qua* thinker to be achieved, including the development of self and character, the acquisition and confirmation of knowledge, and the development and exercise of moral agency.

h. *Appropriate recognition and treatment*. Each thinker has a fundamental interest in being recognized by other agents for the person she is and having others treat her morally well. (Shiffrin 2014, 86-88)

To Shiffrin (2014), the realization of these interests depends upon the ability to convey our thoughts freely to others. Or, as she puts it, "speech and expression are the only precise avenues by which one can be known *as the individual one i*s by others. If what makes one a distinctive individual *qua person* is *largely* a matter of the contents of one's mind, to be known by others requires the ability to transmit the contents of one's mind to others" (Shiffrin 2014, 89). What is of fundamental interest to us is to be known as the individuals we are, and to accomplish that we need to be fully respected by others (Shiffrin 2014). To develop ourselves as thinkers and to form true beliefs, we need some external input, namely we need others' reactions and responses to our beliefs (Shiffrin 2014). For this, undisrupted communication is crucial, i.e., free speech is crucial. What follows is that one could question whether restricting hate speech would be consistent with this kind of argumentation. I will leave the answer to that for later after we have established the kind of harm hate speech produces and have fleshed out the definition of hate speech.

1.3. Harm or offense

In order to build a rationale for possibly restricting hate speech, one must first make a clear distinction between harm and offense. As we know, hate speech is indeed offensive. But, as Simpson (2013) notes:

> We can take it as a given that hate speech is morally benighted and (often) profoundly offensive. But if that is *all* we can say about the adverse character and effects of hate speech, the other putative rationales for its restrictions seem relatively tenuous. We may allow (i) that it is *wrong* to behave rudely, (ii) that reactionary ideas *deserve* a repri-mand, (iii) that we have good reasons to formally *express* our opposition to gratuitously offensive speech, (iv) that we have good reasons to *compensate* hurt feelings, (v) that we ought to fairly *distribute* the burdens of incivility, or (vi) that persistent, gratuitous offensiveness can be a kind of *oppression* that people would rightly want to avoid. But at most, these things seem like defeasible *pro tanto* reasons for restricting hate speech – reasons liable to be overridden by considerations that countervail against legal restrictions on *any* conduct (e.g. costliness, risk of inefficacy, risk of sinister misuse), quite apart from the countervailing free speech concerns that specially apply in this arena. (Simpson 2013, 4-5)

In light of that, argues Simpson (2013), if we want to place restrictions on free speech, the best way forward is to establish that hate speech harms its targets.

Some authors, such as Feinberg, feel that even offense would be enough to restrict hate speech, but others, such as Jeremy Waldron, stress that mere offense would not warrant restrictions. In the following text, I will provide some insight into both of these accounts.

Feinberg (1985) proposes what he calls the offense principle when it comes to regulating speech that states:

> It *is always a good reason in support of a proposed criminal prohibition that it would probably be an effective way of preventing serious offense (as opposed to injury or harm) to persons other than the actor, and that it is probably a necessary means to that end* (i.e., there is probably no other means that is equally effective at no greater cost to other values). The principle asserts, in effect, that the prevention of offensive conduct *is* properly the state's business. (Feinberg 1985, 1)

Feinberg understands the word “offend” to mean: “to cause another to experience a mental state of a universally disliked kind (e.g., disgust, shame)” (Feinberg 1985, 2).

To be offended in the strict and narrow sense one "'must suffer a disliked taste', then 'attribute that state to the wrongful conduct of another' and finally, one must 'resent the other for his role in causing me to be in the state'" (Feinberg 1985, 2). However, Feinberg stresses that the offense principle requires only that one suffers a disliked taste and that the state is wrongfully produced by another party. In other words, the offense principle requires that "there be a wrong, but not that the victim feel wronged" (Feinberg 1985, 2). Since Feinberg views the offense principle in a broader sense, he emphasizes that offense will surely be less serious than harm. Thus, the law should not treat mere offense as serious as certain harms. For example, states Feinberg, a mere fine for an offense should suffice. Feinberg recognizes that people may take offense in various situations. For example, one may take offense because she holds certain prejudices, so seeing an interracial couple might make her feel shocked and disgusted. Because of that, Feinberg states, the offense principle must be very precise in its formulation. In light of that, Feinberg proposes the way in which to weigh the seriousness of a certain offense as follows: the magnitude of the offense, including intensity, duration, and the extent of the offense, the ability to avoid the offense, whether the offense was voluntarily incurred, and whether the person is prone to being easily offended (Feinberg 1985, 35). It is upon legislators to weigh these categories, bearing in mind the social value of particular speech (for example, voicing opinions about public policies).

To my mind, the issue here is that hate speech undoubtedly causes offense, but it *also* causes harm. As Feinberg notices, various people may be offended by various things. That is why building a rationale for restricting speech on offense is a slippery slope.

In *Harm to Others* (1984), Feinberg also examines the issue of harm. He states that the meaning of harm that should be of interest should not be harm done to things, but the meaning should apply to persons and their interests. To Feinberg, the definition of harm that matters in the legal sense is the one that thwarts a person's interests. He distinguishes that definition of harm (a set-back of a person's interests) from a definition of harm that he sees as wrongs, i.e., to say somebody has harmed us is like saying they have wronged us. People can wrong us when they invade our interests, but, he adds, not all invasions of interests are harms (for example, freely consenting to an invasion of our interests would not classify as harm) (Feinberg 1984). The person's interests would, according to Feinberg, be anything one has a stake in, and they are linked to our well-being. Feinberg provides us with examples: a person has an interest in her physical health, emotional stability, good social environment, financial security, freedom from coercion, the absence of pain, and so on. These are welfare interests and according to Feinberg these interests need protecting because they are the most important interests a person has, i.e., they are required for one's well-being. If these interests are in any way thwarted or invaded, then we can say a person has been harmed. According to Feinberg, the harm

principle overlaps the definition of harm as a set-back of interests and a wrong but claims that the harm principle is of little help to legislators when it comes to legal issues.

My thoughts on the issue, as I have stated, are that harm is a more appropriate notion than offense when referring to the effects of hate speech. Offense, as Jeremy Waldron argues in his book *Harm in Hate Speech* (2012), is a subjective reaction and, as per the dictionaries, it refers to hurting one's feelings and, as such, would not warrant protection. For Waldron, what warrants protection is an attack on a person's dignity, where dignity is understood as one's status and basic social standing as being equal to other members of society and having certain entitlements, such as basic rights. Furthermore, leaning on Rawls, he argues that in a well-ordered society, members rely on one another to uphold justice, but if the majority uses hateful speech to refer to minorities, then it means that the majority doesn't recognize nor respect the minorities' social standing. This, in turn, means that minorities cannot be assured that the majority will uphold justice (Waldron 2012).

I side with Waldron in arguing that offense, being a subjective reaction, would not warrant protection. Hate speech undoubtedly offends, but I see a problem with this notion. There can be cases where people are not even offended by vicious things said to them or about them. However, the fact that they did not take offense does not take away from the harm being done, whether they are aware of it or not.[7] For example, historically marginalized individuals who have been targets of hate speech may adapt to systematically being treated as lesser in society and thus may accept that as the norm when, in fact, it shouldn't be. The problem with hate speech, as we will see, is that it can cause harm that can have a primary and a secondary effect, which is tied to prejudice. The primary effect will be the degradation of the target on the spot, while the secondary effect will manifest in the long term. These effects of hate speech will be the core issues in this thesis, which I will elaborate on in the forthcoming chapters. Examining these effects will provide reasons for restricting hate speech, as well as introducing a novel notion of an injustice that I refer to as derogatory-labeling injustice, which will help us understand the effect of hate speech.

[7] Further elaboration on this will be provided in Chapter IV.

1.4. What is hate speech?

1.4.1. Defining hate speech

Hate speech,[8] and consequently, freedom of speech, has been a long-debated subject in various fields of study, and for a good reason. As I have pointed out in the Introduction, a spoken word can be a powerful tool. Let us remember that in the Middle Ages, women labeled a witch were persecuted. In order to save himself from a similar destiny, Galileo Galilei had to renounce his words because they went against the current teachings of the Church. All these examples show just how powerful words can be. The debate on hate speech is not an easy one. On the one hand, we have a speech that can probably cause certain harms to individuals and groups of people and, on the other, we want to protect freedom of speech since it is one of the fundamental rights in liberal societies, and infringing on that right has to be based on very good reasons. As we have seen, there are sound arguments on both sides of the debate. Then, I will try to present a rationale that is, in my opinion, suitable for restricting hate speech. But in order to get to that point, there are some steps we need to take. First of all, we need to understand the current issues with hate speech. One of the first problems we need to address is that there is no consensus about the definition of hate speech. This is important because definitions vary and, consequently, the legal treatment of hate speech differs greatly across countries. So, there is no theoretical nor legal consensus on what hate speech is, meaning that different countries have different regulations of hate speech. The first task will thus be to list some of the most known available theoretical definitions of hate speech and try to pinpoint their joint characteristics to get an idea of what constitutes hate speech. Then, in order to understand why it is important to try to find a satisfactory definition of hate speech and how the absence of one causes confusion, not only in the theoretical sense but, consequently, in the legal sense as well, we will examine how different countries treat hate speech. For this task, I will concentrate on two most known and most evident differences in the legal treatment of hate speech, and that is the one between the US and Europe, where the US has generally more liberal laws leaning on the teachings of John Stuart Mill, and European laws which are, to put it plainly, more strict. This will help us understand why hate speech remains to be a pressing matter.

Let us begin by examining some of the known and available definitions of hate speech. There are numerous definitions of what might constitute hate speech. For

[8] Usually, the discourse about hate speech is in the domain of public speech, both verbal and non-verbal (written and spoken word, as well as symbols, photos, pictures, etc.). This is the domain that will be accepted in this thesis as well.

example, Matsuda (1989) treats racist speech as a *sui generis* category, focusing on the historical perpetuation of violence as well as the degradation of disenfranchised groups. In addition, she offers categories that can help identify hate speech: to qualify as hate speech, a given speech has to be persecutorial and degrading, directed at a historically disenfranchised group, and based on racial inferiority (Matsuda 1989, 2357). Laura Lederer and Richard Delgado offer a similar account of racist speech that ridicules or negatively portrays historically oppressed people based on who they are (Lederer and Delgado 1995). Samuel Walker also focuses on victimized groups that are targets of any form of offensive speech based on race, ethnicity, religion, or nationality (Walker 1994).

Similar to the aforementioned definitions, various documents incorporate some form of definition of hate speech.

For example, The Council of Europe has issued a recommendation No. R (97) 20 of the Committee of Ministers to member states on hate speech. In it, they condemn and urge member states to take action against "all forms of expression which incite racial hatred, xenophobia, anti-Semitism and all forms of intolerance, since they undermine democratic security, cultural cohesion and pluralism" while "noting that such forms of expression may have a greater and more damaging impact when disseminated through the media" (Council of Europe, *Recommendation No. R (97) 20*, 106).

The European Convention on Human Rights guarantees freedom of expression, but not without limits. These freedoms can be subject to restrictions "in the interests of national security, territorial integrity or public safety, for the prevention of disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others, for preventing the disclosure of information received in confidence, or for maintaining the authority and impartiality of the judiciary" (European Convention on Human Rights, *Article 10*, 1950, 12).

In 1966, the United Nations adopted the International Covenant on Civil and Political Rights where article 19 of the Covenant protects freedom of expression:

> "Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice." But, the following article, article 20 states that "any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law."

In addition to that, the United Nations also adopted the International Convention on the Elimination of All Forms of Racial Discrimination which states that "all propaganda and all organizations which are based on ideas or theories of superiority of

one race or group of persons of one color or ethnic origin, or which attempt to promote or justify racial hatred and discrimination in any form" should be prohibited.

We can thus conclude that a unified definition of hate speech does not indeed exist, but by dissecting some of the definitions and codes mentioned, we can single out joint characteristics. Accordingly, it seems that most definitions and recommendations about hate speech refer to:

a) people or groups of people who have certain ascribed characteristics, and,

b) its focus on public speech that can be disseminated. This perlocutionary effect seems important because it appears that the focus is on "spreading" hatred.

These characteristics of hate speech will be good enough for now.

As mentioned before, the implementation of these theoretical definitions and recommendations varies across countries. Not having a unified definition of hate speech pours into the legal domain, which leads to each country having a different stance on hate speech. It is thus useful to look at how various countries treat hate speech, and which hate speech laws they implement.

1.4.2. Legal treatment of hate speech in different countries

Brown (2015) offers a detailed review of various treatments of hate speech laws in various countries.[9]

He mentions ten clusters of laws, regulations, or codes that are used to limit hate speech. The first cluster would be group defamation, which Brown explains as defamation of members of a group based on their ascribed characteristics. He differentiates between two kinds of laws, catchall and *sensu stricto,* wherein defining the former, he relies on the distinction made by David Riesman (1942), who explains that catchall can encompass laws dealing with expressions of falsehoods towards a certain group, laws which can incite hatred towards or between groups, and laws that can incite a breach of peace. Such laws can be found in the US, the UK, and mainland Europe. The *sensu stricto* laws are concerned with false statements about groups of people and can damage the reputation of a group. These laws can be found in the Netherlands, Slovakia, Spain, Germany, Israel, even in parts of the US, and in the media law in countries such as France and Ivory Coast. The second cluster, according to Brown (2015), concerns negative stereotyping or stigmatization that is mainly found in laws concerning the media, where restrictions are imposed on stereotyping based on ascribed characteristics. Brown

[9] For detailed analysis see Brown (2015).

provides an example of a BBC Radio 1 presenter who made a joke about a person's sexual orientation on air and was in breach of the Broadcasting Code because of that. The third cluster centers on the expression of hatred. Again, the focus is on groups with certain ascriptive characteristics, and these codes/laws impose restrictions on insulting and slurring directed at such persons and can be found in Belgium, Bolivia, Cuba, Croatia, Denmark, Ecuador, Greece, Indonesia, Italy, Norway, Rwanda, Sweden, Turkey, parts of the UK, and the state of Connecticut. The fourth cluster imposes restrictions on any utterance that can incite hatred towards a group with ascribed characteristics, and such laws can be found in all parts of the world. The fifth cluster centers on threats to public order, which prohibits speech directed at groups with ascribed characteristics, as those threats are likely to endanger public safety. These laws can be found in many countries, such as Canada, Egypt, Germany, India, Turkey, etc. The sixth cluster focuses on instances of expressions that deny, trivialize, or even glorify acts of mass cruelty, violence, or genocide, such as the Holocaust. These laws can be found in, e.g., Canada, Czechia, Israel, Romania, Spain, Switzerland, France, Germany, Austria, or Belgium. The seventh cluster of laws is concerned with dignitary crimes or torts, i.e., humiliating or degrading people with ascribed characteristics, and these laws can be found in Canada, Costa Rica, Germany, Switzerland, or China, and, in some countries, it takes the form of face-to-face interaction, such as in Brazil, South Africa, and, in some cases, the US. The eight cluster centers on violations of civil or human rights and is mostly focused on anti-discrimination laws. These laws can be found in the US law, and, also, some instances of this cluster of laws/codes are incorporated in some campus speech codes. Countries such as Canada, New Zealand, and the UK have also incorporated this cluster into their laws. Expression-oriented hate crimes are the focus of the ninth cluster, and the laws are focused on penalizing speech fueled by hatred of people with ascribed characteristics and, as such, can be found in, e.g., Croatia, France, Italy, Russia, the UK, and the US. The last cluster of regulations cannot be considered hate speech law in the narrow sense because these laws restrict the use of hate speech by imposing time, place, and manner restrictions. For example, various US states have regulations against protests at funerals. Finally, Brown emphasizes the three jurisdictional levels of these laws:

> ...laws/regulations/codes at the level of a sovereign state, including laws issued by national, state, county, city, and even village governmental authorities (e.g., constitutional law, criminal law, civil law, administrative law, immigration law, public gatherings law, law on contempt of court, various kinds of local or municipal laws, codes, regulations, or ordinances); laws/regulations/codes at the international level (e.g., conventions, declarations, protocols, the jurisprudence of supranational human rights

> courts); laws/regulations/codes at the level of subnational institutions, organizations, and commercial companies (e.g., speech codes enforced by employers, schools, universities; rules on permissible content imposed by independent media and internet regulators; standards or codes on acceptable content adopted and enforced by newspapers, TV and radio broadcasters, internet service providers, social networking websites, internet messaging services). So in that sense one can plausibly say that as a corpus of law hate speech law is not merely variegated in form but also broad in scope. (Brown 2015, 39-40)

We shall now turn our focus to the different treatment of hate speech in different countries, and the biggest difference can probably be seen in the European codes/laws and the codes/laws in the US. To pinpoint those differences, I will use laws that exist in the US and the laws that exist in Croatia since Croatia, as do most of the European countries, follows the recommendations of the Council of Europe.

Hate speech is protected in the US under the First Amendment that states: "Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the Government for a redress of grievances".

Courts in the US have been very protective of freedom of expression, including hate speech. However, there is one exception to this, and that is in the case of fighting words. In Chaplinsky vs. New Hampshire[10] in 1942, the Supreme Court stated that:

> there are certain well-defined and narrowly limited classes of speech, the prevention and punishment of which have never been thought to raise any Constitutional problem. These include the lewd and obscene, the profane, the libelous, and the insulting or "fighting" words – those which by their very utterance, inflict injury or tend to incite an immediate breach of the peace. It has been well observed that such utterances are no essential part of any exposition of ideas, and are of such slight social value as a step to truth that any benefit that may be derived from them is clearly outweighed by the social interest in order and morality. (Chaplinsky vs. New Hampshire 1942)

As such, fighting words are not protected by the First Amendment. It is evident that the doctrine of fighting words is very narrow, and the focus is on face-to-face

[10] For further information see Chaplinsky v. New Hampshire.

confrontation, which has to constitute a clear and present danger, or as Stephen W. Gard explained:

> This ambiguous passage suggests three rationales in addition to the prevention of responsive violence to justify the censorship of fighting words: (1) that such words are not "speech" within the meaning of the first amendment because they are unnecessary to the expression of ideas and thus lack social utility; (2) that such words are akin to verbal assaults and inflict emotional distress upon their recipient; and (3) that whatever slight social value such words may possess is per se outweighed by the psychic injury and responsive violence caused by them. (Gard 1980, 534)

The Supreme Court of the United States has also listed the elements necessary for an utterance to constitute fighting words. Gard (1980) summarizes:

> In addition to the requirement of intent, common to all speech crimes, four elements must be present before the doctrine will deprive a message of constitutional protection. First, the utterance must constitute an extremely provocative personal insult, a factor requiring a judicial analysis of the content of the expression. Second, the words must have a direct tendency to cause an immediate violent response by the average recipient. Third, the words must be uttered face-to-face to the addressee. Fourth, the utterance must be directed to an individual, not a group. These final three requirements are contextual in nature and mandate a judicial evaluation of the circumstances in which the speech is uttered. If any of these four elements is absent, the expression may not be denied constitutional protection on the ground that it comes within the scope of the fighting words doctrine. (Gard 1980, 536-537)

However, ever since Chaplinsky v New Hampshire, the Supreme Court has been reluctant to apply the fighting words to other cases. To illustrate this point, I will provide some example cases where hate speech was protected under the First Amendment and where hate speech did not fall under unprotected fighting words.

One of those cases is Collin v Smith, a case from 1978. The Nationalist Socialist Party of America planned a rally wearing Nazi symbols such as swastikas on their uniforms in the village of Skokie, which had a significant Jewish population, some of them even Holocaust survivors. The villagers wanted to protect themselves from what they considered to be a traumatic experience and filed an injunction. At first, and at the lower courts, the injunction was granted, and the Party was not allowed to march and to wear symbols such as swastika. This ruling was appealed, and the Party was later allowed to march but without swastikas. However, through appeals, the case reached the Supreme

Court of the United States, which then ruled that the Party could march, that the swastika was protected under the First Amendment, and that it does not constitute fighting words.

The other two cases are similar in that they involve cross burning. Cross burning is a symbol of intimidation against African Americans, used usually by the Ku Klux Klan organization. The first case is the 1992 case R.A.V. v St. Paul, where a group of teenagers burned a cross in the backyard of an African American family. The decision of the lower to sanction the action as a misdemeanor reached the Supreme Court, which reached the decision that cross burning cannot be regulated and prohibited based on its content and is thus protected by the First Amendment. The second case is a more recent, 2003 Virginia v. Black case, where the Court held that some cases of cross burning can be considered a true threat and thus unprotected by the First Amendment. Still, the Court also held that cross burning in itself does not constitute an intent to intimidate. In other words, unless proof of an intent to intimidate was found, cross burning is protected under the First Amendment.

It is clear from the abovementioned cases that the fighting words doctrine is narrow and that the interpretation of what constitutes fighting words varies from court to court. Thus, a lower court may prohibit an utterance it considers to be unprotected, and the higher courts may overrule such decisions as unconstitutional. Similarly, some scholars argue for a broader definition of fighting words that would broaden the scope of possible prohibitions on utterances, and some scholars argue for an even narrower definition or even a complete abandonment of the doctrine.

Let me now shift the focus from the US to Europe or, more precisely, to Croatia's legal treatment of hate speech.

In its treatment of hate speech in legal terms, Croatia follows the recommendation of the Council of Europe, and these recommendations are usually far less forgiving when it comes to hate speech than US laws.

Firstly, the Croatian Constitution guarantees all rights and freedom regardless of race, religion, ethnicity, sex, language, social status, education, and so on. Article 16 of the Constitution allows some restrictions in order to protect personal or social values. Article 39 states that the use of violence or spreading hatred based on nationality, race, religion, or any other form of intolerance can be sanctioned.

Furthermore, Article 325 from the Croatian Criminal Code states that:

(1) Whoever in print, through radio, television, computer system or network, at a public gathering or in some other way publicly incites to or makes available to the public tracts, pictures or other material instigating violence or hatred directed against a group of persons or a member of such a group on account of their race, religion, national or ethnic origin, descent, colour,

gender, sexual orientation, gender identity, disability or any other characteristics shall be punished by imprisonment not exceeding three years.

(2) The same punishment as referred to in paragraph 1 of this Article shall be inflicted on whoever publicly approves of, denies or grossly trivialises the crimes of genocide, crimes of aggression, crimes against humanity or war crimes, directed against a group of persons or a member of such a group on account of their race, religion, national or ethnic origin, descent or colour in a manner likely to incite to violence or hatred against such a group or a member of such a group. (Croatian Criminal Code, Art. 325)

In addition to Article 325, Article 147 of the Criminal Code refers to insulting others and states that whoever insults another can be punished by a fine, even more so if the insult is uttered through the media. The Code, however, does not specify what institutes an insult.

In addition to the Croatian Criminal Code, there are a variety of other acts that focus on preventing hate speech. The Media Act, for example, forbids media "to support and glorify national, racial, religious, sexual or other discrimination or discrimination based on sexual orientation, ideological and national entities and encourage national, racial, religious, sexual or other hostility or intolerance, hostility or intolerance based upon a sexual orientation, violence and war" (Media Act, Art. 3).

Croatia's Anti-discrimination Act in Article 25 of the Act also predicts penalty provisions for discrimination as follows:

(1) Whoever, with the aim to intimidate another person or to create a hostile, degrading or offensive environment on the grounds of a difference in race, ethnic affiliation, color, gender, language, religion, political or other belief, national or social origin, property, trade union membership, social status, marital or family status, age, health condition, disability, genetic origin, native identity or expression, and sexual orientation, hurts another person's dignity, shall be charged a fine for misdemeanor amounting from HRK 5,000.00 to HRK 30,000.00.

I will mention one more law, which is a Law on Misdemeanors against Public Order and Safety that in Article 5 says the following:

Whoever disturbs public order and peace by performing, reproducing songs, compositions, and lyrics, or carrying and displaying symbols, texts, pictures, or drawings in a public place shall be punished for a misdemeanor by a fine in the local currency equivalent of 50 to 300 DEM or imprisonment for up to 30 days. Whoever distributes printed or recorded things in an unusually intrusive or impudent manner, thereby

disturbing the peace of citizens, shall be punished for the offense by a fine in the equivalent of domestic currency of 50 to 300 DEM or imprisonment for up to 30 days.

Now that I have presented some existing definitions of hate speech and juxtaposed hate speech laws in the USA and Croatia, I will briefly tackle the execution of said laws. Briefly, because I am inclined to leave in-depth analysis to law experts and will only give my insight and intuitions about the aforementioned definitions and laws.

We have seen that, in theoretical discourse, there is still no unified definition of hate speech, although there are similarities. I have fleshed them out earlier in the text. As far as the law is concerned, Vesna Alaburić (2003) notes that despite the lack of a theoretically unified definition, we can still conclude that hate speech has a relatively precise meaning.[11] Since there is no theoretically unified definition of hate speech, each country is responsible for prescribing laws concerned with hate speech, and that's the reason why hate speech laws may differ from country to country, as we have seen in the example of the USA and Croatia. But, even though hate speech is relatively more precisely defined where the law is concerned, I still feel that there is room for improvement. As I have pointed out earlier, there are cases in the US where the lower and higher courts will not agree on what constitutes fighting words. I have provided examples where there are different interpretations of the fighting words doctrines, which has led to different verdicts for the same case. In Croatia, there have been cases where, on separate occasions, the same act has not been treated equally. At one time, the act was prosecuted, and at another time, it was not.[12] This, however, has more to do with law enforcement and application. Nevertheless, that is not to say that laws concerning hate speech in Croatia do not need more work and even more precise definitions. Here I will briefly try to explain my position.

I have already pointed out some problems that arise in practice in the US concerning the fighting words doctrine, but now I would like to focus on Croatia's laws mentioned in the above text. Firstly, Article 325 from the Croatian Criminal Code states that "Whoever in print, through radio, television, computer system or network, at a public gathering or in some other way…" etc. The potential problem I see here is the phrase *or in some other way*. This notion is pretty unclear and can be interpreted in various forms. Also, Article 147 of the Criminal Code mentions insulting the other, but the article does not define what institutes an insult. This is also a potential problem because these unclear or rather vague definitions leave room for subjective interpretation of the matter, which potentially leads to uneven and different treatment of hate speech even in the same country (in our example, Croatia). I detect the same problem with the Article 25 of the

[11] A thanks goes to Professor Sanja Barić who also pointed me in this direction.

[12] The case I have in mind is the verdict for the saying "Za dom spremni".

Anti-discrimination Act where it is stated that "Whoever, with the aim to intimidate another person or to create a hostile, degrading or offensive environment…" etc. The potential issue I suspect here is the phrase *offensive environment*. Again, it is unclear what would be offensive. To be offended is a matter of feeling, it is very subjective, and people can be offended by various things. For example, one can be offended by, say, another person showing too much skin for their taste if they come from a conservative background. Or one may not be offended at all when a speaker clearly insults them, let's say, by using a racial slur. I suspect these issues could potentially cause confusion in law enforcement and that it leaves too much room for subjective interpretation of the law. It would surely be more helpful if the laws were written more clearly, i.e., if they were more specific.[13] However, due to the scope of this work, I do not wish to untangle these issues here and shall leave this for some future occasion. My aim was to pinpoint potential issues that may arise from laws written in such a manner. To add, this does not apply only to Croatia because we have seen from the examples in the USA that the same problems arise there with defining what falls under fighting words. This is why the debate about hate speech and about what and how, and even if, hate speech should be sanctioned is, to figuratively put it, a hot potato.

To reiterate, there is still no unified definition of hate speech, although some similar points can be found in most definitions and codes. Various countries treat hate speech differently, so laws about hate speech vary greatly; some, like the US laws, are more liberal, and some, like the EU laws/codes, are less forgiving when it comes to the legal treatment of hate speech. The question is—which approach to hate speech should one then take? Answering this burning question is no easy feat. There are a number of things that should be taken into account when discussing it. What reasons warrant restricting freedom of speech that is so important in liberal societies? Furthermore, is restricting freedom of speech the only approach one can take when encountering hate speech? In the following Chapters, I will focus my attention on one particular aspect of hate speech—slurs. I will delve into the philosophy of language to explain what slurs are and, most importantly, what they do when uttered. Then, I will shift my focus to the potential harm slurs can do to individuals and groups, where I will help myself with some empirical evidence on stereotypes. This examination will yield a new kind of injustice that arises from the use of slurs—derogatory-labeling injustice. After that, I will pinpoint some ways in which we can counter such speech.

[13] However, there are some who would not agree with this and that would argue that the demand philosophers put on the legal domain is too high.

CHAPTER II: BUILDING THE NEEDED BACKGROUND

2.1. Introduction

In the previous Chapter, I introduced the debate about freedom of speech and presented some of the most known arguments for protecting free speech and arguments for restricting speech. I also gave an overview of how hate speech laws work in practice and pointed out some drawbacks I found with the current laws, where I concentrated on Croatia's hate speech laws. Since my main claim in this thesis is that slurs evoke a specific kind of prejudice, the negative identity prejudice introduced by Fricker (2007), some background on socialization, stereotypes and prejudices is needed. To reiterate, the stereotype that is the driving force of slurs is a specific one—the identity prejudice proposed by Fricker. Namely, I borrow from Miranda Fricker's work on testimonial injustice, where she identified a specific kind of prejudice at work in testimonial injustice, and I will claim that the same kind of prejudice is evoked when uttering a slur. But, before making this move, I have to set a needed background. First, in order to see what kind of effect slurs may have on us, we have to examine how socialization may shape us as individuals since a large part of socialization is language. Language is the most important tool for communicating, and communication is crucial for being a part of a community. Secondly, in my view, prejudices are evoked by slurs, and I will present empirical evidence on how stereotypes and prejudice may affect us. This will set a basis for the claim that slurs harm us and that the harm derives from prejudice evoked by systematic uses of slurs. Finally, since I will borrow from Miranda Fricker's work on testimonial injustice, I will give a brief summary of her work on testimonial injustice, which I will later transfer to the issue of slurs.

2.2. Socialization

Before going further into this investigation, I will briefly explain what socialization is and how it works because this is important for the task at hand.

Socialization is defined as "processes whereby naive individuals are taught the skills, behavior patterns, values, and motivations needed for competent functioning in the culture in which the child is growing up. Paramount among these are the social skills, social understandings, and emotional maturity needed for interaction with other individuals to fit in with the functioning of social dyads and larger groups" (Maccoby 2007, 13). When a child is born, the first socialization it encounters is with the parents and caregivers. Later, this circle broadens to teachers and peers, and the process of socialization continues throughout one's life.

Socialization has been a long-studied process and has seen many theories, from Freud's stages of personality development, Piaget's four stages of cognitive development, Kohlberg's stages of moral development, to Erikson's eight stages of development. For a long time, the relationship between parents and children has been in focus when it comes to socialization. In recent decades, the focus has widened to genetics, also reaching to the role different media play in that development. Here, my focus will be on the role of socialization in our development since this is important for our task.

Various authors have stressed the importance the environment plays in our development, with one of them being Kevin W. Saunders, who begins by mentioning Virgil's saying, "As the twig is bent, the tree inclines." (cited in Saunders 2011, 168), emphasizing how we are formed during a young age. There is plenty of research showing how the environment can affect us, and Saunders stresses one in particular that showed the salient role the amygdala can play in recognizing facial expressions where we, through experience, learn to assign labels to specific facial expressions (A. Baird et al. 1999).

Robert Post, in his paper titled simply *Hate Speech* (2009), stressed how social norms are crucial to one's identity because they become internalized, and he quotes what he believes to be the best fitting description of this process proposed by George Herbert Mead, which says:

> What goes to make up the organized self is the organization of the attitudes which are common to the group. A person is a personality because he belongs to a community, because he takes over the institutions of that community into his own conduct. He takes its language as a medium by which he gets his personality, and then through a process of taking the different roles that all the others furnish he comes to get the attitude of the members of the community. Such, in a certain sense, is the structure of a

> man's personality . . . The structure, then, on which the self is built is this response which is common to all, for one has to be a member of a community to be a self. (Mead 1962, 162)

One possible example of this is the expectations that society has concerning our gender, with one of them being an expectation to act according to one's gender (Ning, Dai, and Zhang 2010). It is useful here to distinguish between sex and gender, as sex generally refers to biology, and gender is considered to be a social construct or, in other words, it is socially and culturally determined. One of the most famous quotes about the distinction between sex and gender is given by Simone de Beauvoir (1949), where she states that you are not born as a woman, but you become one, stressing the importance gender roles play in our life and how social roles determine our identity as a woman or as a man.

Another author concerned about how the environment might shape us as individuals is Andres Moles, who proposes the idea of mental contamination, which is defined as "the process whereby a person has an unwanted response because of mental processing that is unconscious or incontrollable" (as cited in Moles 2007, 69). According to Moles, this presents a threat to one's autonomy, and he gives three reasons why that is the case. First, "it presents an obvious challenge to the condition of identification in so far as agents whose responses are contaminated cannot identify with them" (69-70). Second, "contaminated responses are not the outcome of reasons but of external influences that have not been considered" (Moles 2007, 70), and third, "mental contamination poses an important threat to our critical judgment, to the extent that it makes us react in ways we would not want to" (ibid.).[14] Moles (2007) mentions that "among other features and mechanisms that trigger the automatic response, we have to consider the society which, in this case, creates social stereotypes and triggers unwanted responses" (Perhat 2016, 233). Some people's belief systems may correspond to these responses, and if that is the case, then their autonomy is not violated. However, for people whose belief system does not correspond to social stereotypes, these automatic responses are something to regret (Moles 2007). According to Moles (2007), the best thing to ameliorate this effect would be to simply avoid being exposed to contamination sources, as Wilson and Brekke (1994) have suggested. But, Moles (2007) is aware that this strategy is not free of problems even though we should try "to categorize the weights of different forms of contamination (racial and gender based are particularly important)" (74). As mentioned in my previous paper, "Moles thinks that this strategy is the best one because he thinks that any rational discussion about social stereotypes that would potentially lead to understanding that these stereotypes are false will not have any effect because 'many people who believe they are not racists still manifest racist reactions'" (Moles, 2007, p. 73)" (Perhat 2016, 233).

[14] As cited in Perhat (2016, 233).

There has been some empirical research that may corroborate Moles' intuitions about these unconscious processes. Devine (1989) conducted an experiment where she presented the subjects with subliminal words stereotypical of African Americans. Those in the experimental group were presented with many words that negatively describe African Americans as opposed to the control group. The subjects then had to read a story about a man with ambiguous behavior and rate his character traits as hostile or non-hostile. The results showed that the experiment group rated him as significantly more hostile than the control group. Devine's (1989) explanation of this automatic stereotyping was that subjects were brought up in an environment (the USA) where everyone has some knowledge of stereotypes associated with race, and when coming into contact with a certain race, these stereotypes will be activated (Brown 2010).

Brown (2010) notes that similar findings about automatic stereotypical associations were made by Correll and colleagues (2002), where participants had to make a quick decision about what kind of an object was held by a video game character who was sometimes black and sometimes white. More participants concluded that the object the person was carrying was a gun when the game character was black. Moreover, when the participants had to decide whether to shoot an armed target, they were quicker to shoot at black targets, and when the game character was not armed, the participants "took longer to decide not to shoot at a black target than at a white target" (Brown 2010, 87).

Furthermore, research by Amodio et al. (2004) on automatic prejudiced responses suggests that people who don't hold prejudiced beliefs tend to recognize the need to implement control over unintentional bias when it occurs but fail to implement the intended response.

As presented above, the process of socialization is essential for understanding stereotypes and prejudice. Through socialization, we learn the norms about culture, but we also come into contact with stereotypes and prejudice. Since these are crucial for our task, I will set my focus on explaining how stereotypes and prejudice work. There are numerous studies about how stereotypes and prejudice may affect us, but empirical research on the effect of slurs is scarce.

2.3. Stereotypes and prejudice

According to Devine (1989), stereotypes are established early on, even from 5 years old. Another research (Anzures et al. 2013) showed that newborns can express a preference for their race if that race is all they have seen around them. But what are stereotypes, and how exactly do they work? Stereotypes work as sort of a shortcut our brain creates to make life easier for us. It is easier for us if our brain categorizes things, and we begin to do it as soon as we are born (Cikara and Van Bavel 2014). Not all stereotypes are bad. There are positive stereotypes where we think that a group of people have a certain advantage over others. For example, we often think that black people are more athletic and better at sports such as basketball. Whether or not these positive stereotypes can be viewed as an advantage to the group of people they are referenced to is questionable since not all individuals who are a part of that group will hold the same characteristics. I will not be dealing with positive stereotypes here; I will rather shift my focus to negative stereotypes since they are at the core of our interests. It is clear that stereotypes are formed early on and that the experiences we have can shape our stereotypes. So, what is a stereotype? The first one to use the word was Walter Lippmann (1922), and he described stereotypes as pictures we carry in our heads. We can describe a stereotype as "a generalization about a group of people in which identical characteristics are assigned to virtually all members of the group, regardless of actual variation among the members" (Aronson et al. 2015, 416). Stereotypes can linger on in our minds and affect our judgment even if we are not fully aware of it (Fricker 2007). Fricker, whose theory of testimonial injustice I will discuss in more detail later, describes stereotypes as "widely held associations between a given social group and one or more attributes" (Fricker 2007, 30), and according to this, "stereotyping entails a cognitive commitment to some empirical generalization about a given social group (women are intuitive)" (Fricker 2007, 31).

Prejudice, as Samson (1999) explains, "involves an unjustified, usually negative attitude towards others because of their social category or group membership" (Samson 1999, 4).

As mentioned before, studies of how stereotypes and prejudice affect us are in abundance, and it is impossible to review all of them. Thus, I will focus only on the most prominent empirical research throughout the years. One of the earliest studies of ethnic and national stereotypes, as Rupert Brown (2010) notes, was conducted by Katz and Braly (1933), who found out that, as well as some changes, there was significant stability over the years in the endorsement of group stereotypes. This study actually supports the claim that stereotypes have a socio-cultural origin (Brown 2010). There are, of course, other explanations of the origin of stereotypes, such as the grain of truth theory, which claims that we tend to exaggerate attributes of a given group, which then become stereotypes over time

(Brown 2010). Another possible origin of stereotypes, as Devine and Sherman (1992) note, is their ideological function in that referring to the underprivileged minority group as lazy helps to rationalize the social system (Brown 2010). Or, as Hamilton and Gifford (1976) pointed out, stereotypes may result from the fact that we tend to have better memory of things that happen infrequently. One more possibility of the origin of stereotypes comes from the concept of entitativity. For example, ethnicity or gender have a so-called intermediate entitativity, i.e., the degree to which they are perceived as a unit. The higher the entitativity of a group, the better the possibility of being perceived as a group and thus more easily stereotyped (Brown 2010).

Assuredly, one of the most prominent studies of the effect of bias and stereotypes was conducted by Maass and colleagues (1989) on linguistic intergroup bias. The research showed that an out-group member who presented an undesirable behavior was described in an abstract way (for example, being aggressive), as well as an in-group member who was engaged in a desirable behavior (for example, being helpful). Abstractedness is linked to generalizations about the target. These findings imply that "linguistically abstract communications are perceived as providing more information about the actor than do concrete ones (Experiment 3; see also Semin & Fiedler, 1988a)" (Maass et al. 1989, 990) and suggest that "abstract descriptions are perceived as relatively stable over time (Semin & Fiedler, 1988) and consequently produce the expectation that the (undesirable) action be repeated in the future (Experiment 3)" (ibid.). In other words, the experiment implied that abstract description tends to lead to biased perception, i.e., stereotypes that are then transmitted further.[15]

The previously described experiment by Maass and colleagues shows us how stereotypes may affect our judgment. But there is also empirical evidence of how stereotypes affect their targets, or, as Brown (2010) puts it, stereotypes can be self-fulfilling prophecies. For example, research found that students' performance can be influenced by the teachers' expectations (Rosenthal and Jacobson 1968; Crano and Mellon 1978; Madon and colleagues 2001). An interesting study was done by Eccles-Parsons and colleagues (1982, 1990). The studies showed how parental expectations regarding their children's competencies are influenced by the children's sex, and that, in turn, influences the child's self-perception (Brown 2010). A study done by Harris and colleagues (1992) showed how the belief that their study partner had a hyperactive behavior disorder influenced the perception of the difficulty of the task, where the children said the task was more difficult when they believed their partner had a problem. The children labeled to have such a disorder (whether they really had it or not) have also evaluated the task as difficult, which follows the findings of the self-

[15] For more detailed information on the experiment see Maass, A., Salvi, C., Arcuri, L., & Semin, G.R. (1989). *Language use in intergroup contexts: The linguistic intergroup bias.*

fulfilling prophecies. The children who were not led to believe their study partner had such a disorder did not have such an evaluation.

As mentioned before, stereotypes emerge early, and some studies showed how children, even though several sorting criteria were presented, such as gender, age, and so on, chose ethnicity while sorting photographs of people (Davey 1983; Yee and Brown 1988). However, when the context was changed, for example, when the task was that the photos should be sorted by the "who plays with whom" criterion, gender was predominantly used (Davey 1983). There is also evidence that babies of 3 months prefer to look at members of their race more than members of other races (Kelly et al. 2007). There were other interesting experiments that showed a preference for ethnicity. One of the earliest ones, which was subsequently replicated, showed how children identified themselves with the doll representing their ethnicity (Clark and Clark 1947; Goodman 1952). However, there were instances where black children also preferred the white doll to represent them (Clark and Clark 1947). There were various explanations as to why this happens, and one of them, although inconclusive, was that it was a matter of low self-esteem. Even some anecdotal evidence pointed to that conclusion (for example, the cases where black children would scrub their faces in order to wash the blackness away) (Brown 2010). In terms of gender, it has been established that usually, by the age of 5 or 6, children begin to understand their gender and they prefer their gender for interaction (Brown 2010). In terms of what influences the development of stereotypes and prejudice in children, there are various theories and research that have been done, but according to some research, such as the one done by Castelli et al. (2008, 2009), it was shown that children can pick up non-verbal signals from adults very efficiently so in that respect they can mimic their parents' attitudes (especially their mothers' attitudes, as the research suggested) towards the members of another group. As shown by these studies, stereotypes and prejudice emerge quite early, but the explanation as to why it happens is not so simple. The reason for that is that the research on the impact the parents' attitudes and the mass media have on the development of stereotypes and prejudice in young children is not consistent. Thus, Brown (2010) concludes:

> Confronted with these kinds of problems, social psychologists have developed theoretical models that link the development of prejudice to more general cognitive social and affective changes, which occur in children in the first ten years of their life (Aboud, 1988; Cameron et al., 2001; Katz, 1983; Maccoby and Jacklin, 1987; Nesdale, 2004). Although there are undoubted differences in emphasis between these various theories, the latter have in common the assumption that the child plays a more active role in the developmental process than the traditional socialization explanation allows. In particular, all attribute primary importance to the cognitive capacity for categorization: it both assists

> children to make sense of their environment and provides them with various social identities. (Brown 2010, 136-137)

Brown (2010) explains how Nesdale (2004) gave a possible reason as to why prejudices emerge in older children. Nesdale (2004) attributed the emergence to two factors. One factor is the identification of the child to their ingroup, in which case, if the attitudes of the child's ingroup are negative towards the outgroup, the child is likely to manifest the same negative attitudes. The other factor is the relationship between the ingroup and the outgroup, in which case the more negative relationships tended to produce more prejudice in children.

There is also one more aspect of stereotypes that is useful to consider when examining the possible effects stereotypes can have, and that is the aspect of *stereotype threat*. As defined by Joshua Aronson and Claude Steele (1995), stereotype threat is triggered by a negative stereotype, and its most known effect is to hinder performance. For example, it was found that black students underperformed on a test when it was made known to them that the test would measure intelligence. In contrast, the underperformance was absent when the students believed the test didn't measure anything other than some coping strategies (Steele and Aronson 1995). In addition to underperformance, there are other effects stereotype threat can invoke, such as:

> reduced self-efficacy (Aronson and Inzlict 2004), lowered confidence that one will do well in the stereotyped domain (Stangor et al. 1998); lowered aspirations to pursue stereotype-relevant careers (Davies et al. 2002; Davies et al. 2005); and negative physical and psychological health consequences, including increased general anxiety (Ben-Zeev et al. 2005; Bosson et al. 2004), blood pressure (Blascovich et al. 2001), and feelings of dejection (Keller & Dauenheimer 2003). (Shapiro and Aronson in Stangor and Crandall 2013, 97)[16]

As to why stereotype threat happens, there are various explanations ranging from being concerned about not to confirm the stereotype (Steele and Aronson 1995) to lowering one's expectations, in other words, internalizing the stereotype (Blum 2016). The literature and research about stereotype threat is vast, so it is an impossible task to go into detail about all the mechanisms that go into stereotype threat. Instead, for the conclusion on stereotypes, I will list some existing research and findings about derogatory language and stereotypes.

As we have seen, there are various ways in which stereotypes can affect us both as hearers and as targets. When it comes to the connection between derogatory language and stereotypes, the research that has been done (Bianchi et al. 2019; Fasoli, Maas, and Carnaghi 2015) shows that both neutral and derogatory language, such as slurs, can activate stereotypes, whereas derogatory language tended to trigger a much stronger negative

[16] As cited in Goguen 2016.

evaluative reaction (Cervone, Augoustinos, and Maass 2021). According to Fasoli, Maass, and Carnaghi (2015), in addition to increasing negative attitudes, slurs (specifically homophobic slurs) tend to induce dehumanizing and distancing from gay men. Thus, they conclude that homophobic language may be a way to reinforce and spread homophobia (Fasoli, Maas, and Carnaghi 2015). Cervone, Augoustinos, and Maass (2021) note that five social functions of derogatory language have been identified. One such function is prejudice perpetuation. Similar to the aforementioned research by Bianchi et al. 2019 and Fasoli, Maas et al. 2015, "Bilewicz and Soral (2020) found that digital media users (i.e., the media users more exposed to hate speech) reported higher levels of Islamophobia, a relationship explained by greater acceptance of hate speech" (Cervone, Augoustinos, and Maass 2021, 84). Secondly, derogatory language is linked to the maintenance of status hierarchies. According to Rosette et al. (2013), dominant group members tend to use derogatory language more and are less likely to react defensively to such language. As Cervone, Augoustinos, and Maass (2021) explain, "the asymmetrical use of disparaging language not only reflects the existing social stratification, but it also contributes to maintaining the power differential between dominant and subordinate groups" (84-85). They add: "To quote Simpson (2013, p. 7), 'the best way to make sense of the claim that hate speech in general inflicts harm on people, is to think of hate speech as something that contributes to (identity-based) social hierarchies.' Slurs not only keep social minorities in their subordinate position, but also assure the privileged position of the dominant group" (Cervone, Augoustinos, and Maass 2021, 85). Thirdly, derogatory language can be used to legitimize violence or exclude the outgroup, which can be done by dehumanizing the targets or by portraying them as a threat (Cervone, Augoustinos, and Maass 2021). "The dehumanizing narratives and images have been documented, among others, for homophobia (Fasoli et al., 2016), anti-Semitism (Volpato et al., 2010) and racism (Goff et al., 2008)" (Cervone, Augoustinos, and Maass 2021, 85). In order to portray the target as a threat to the ingroup, there needs to be a strong identification among the members of the ingroup with the ingroup itself, where the members of the ingroup feel that their culture or way of life is threatened by the outgroup and they feel the need to protect themselves (Cervone, Augoustinos, and Maass 2021). Fourth, derogatory language can be used to regulate what is deemed to be socially acceptable behavior among the ingroup (ibid.). For example, derogatory words used for non-heterosexual individuals, which are used in a very broad sense, encourage compliance with male gender norms (Carnaghi et al. 2011). Finally, as Cervone, Augoustinos, and Maass (2021) note, derogatory language also promotes cohesion among the ingroup in the sense that it connects like-minded people (Douglas 2007). In addition to these five social functions, Cervone, Augoustinos, and Maass (2021) also note the negative effect derogatory language has on its victims. For example, they mention diary studies (Swim et al. 2009) with evidence of fear, anxiety, and lower self-esteem experienced by targets of such language. As for the effect on listeners and society, there are several (Cervone, Augoustinos, and Maass 2021). For example, one can

become more accepting of hate speech if one is constantly exposed to it and could even view it as the norm (Soral et al. 2018; Winiewski et al. 2017). Derogatory language can also lead to social distancing from the target group (Winiewski et al. 2017).

All of these aforementioned examples and research show how stereotypes, prejudice, and derogatory language can affect us negatively. Now, I would like to turn to the epistemic issues that can arise from the use of hate speech and slurs. For that task, I will turn to Miranda Fricker's account of epistemic injustice and follow in her footsteps in explaining the harm that can be produced by such language that leans on negative stereotypes.

2.4. Fricker's epistemic injustice: an outline[17]

My aim here is to show how slurs can affect the stakeholders in society and the society as a whole. For this task, I turn to Miranda Fricker, who developed a theory of testimonial injustice in which, among other things, she explains how our social environment can form our identity in her highly influential book *Epistemic Injustice: Power and the Ethics of Knowing* (2007)[18] where she introduced the notion of testimonial and hermeneutical injustice. By presenting Ficker's case for testimonial injustice, I will mirror her argumentation of how slurs can affect us. But, first, to fully understand Fricker's case, we need an outline of her work on testimonial injustice.

To begin with, Fricker needs to set the stage by answering some questions that lead to epistemic injustice. The first point she makes is about power, which she, in the broadest sense, understands as our ability as social agents to influence the social world. She is interested in one particular aspect of power she calls social power, which she describes as "a practically socially situated capacity to control others' actions, where this may be exercised (actively or passively) by particular social agents, or alternatively, it may operate purely structurally" (Fricker 2007, 13). As she further explains, there is one aspect of social power that requires an imaginative social co-ordination, and that is identity power, which she describes as follows:

> There can be operations of social power which are dependent upon agents having shared conceptions of social identity – conceptions alive in the collective social imagination that govern, for instance, what it is or means to be a woman or a man, or what it means to be gay or straight, young or old, and so on. (Fricker 2007, 14)

Thus, identity power presides in the imaginative—in the social conceptions we all share in society. What also presides in these shared social conceptions are stereotypes, and, as Fricker (2007) explains, we use these stereotypes in the testimonial exchange, i.e., when we judge the credibility of the speaker. The problem, as Fricker (2007) notes, is "if the stereotype embodies a prejudice that works against the speaker, then two things follow: there is an epistemic dysfunction in the exchange—the hearer makes an unduly deflated judgment of the speaker's credibility, perhaps missing out on knowledge as a result; and the hearer does something ethically bad—the speaker is wrongfully undermined in her capacity as a knower" (Fricker 2007, 16-17). That is the core of testimonial injustice.

---

[17] This section will heavily lean on my previous work *Pejoratives and testimonial injustice* (2016) published in Miščević, Perhat (eds).

[18] Her work is highly influential, but it also received some criticism that I won't be dealing with here.

Even though, as Fricker (2007) explains, the prejudice at work in the testimonial exchange can result in credibility excess (the speaker receiving more credibility), Fricker is more interested in credibility deficit, or the cases where the speaker receives less credibility because the deficit results in testimonial injustice. According to Fricker (2007), testimonial injustice happens because of the prejudice the hearer harbors. Fricker is interested in the systematic nature of testimonial injustice, so her concern will be a specific kind of prejudice the hearer has—identity prejudice, i.e., prejudice related to our social identity that follows us through every social dimension. Since prejudice is essential for testimonial injustice, Fricker proceeds to explain how the mechanism of prejudice works in testimonial injustice. As she elaborates, prejudice enters our judgment via stereotypes, which she defines as "widely held associations between a given social group and one or more attributes" (Fricker 2007, 30). At times, there can be an identity prejudice in the stereotype, which is common when referring to historically marginalized groups, such as women or people of color. As Fricker (2009) explains, via an example given by Nomy Arpaly (2003), sometimes these judgments can be a non-culpable mistake, as in the case presented by Arpaly (2003) where a boy grew up in an isolated community that believes women to be incapable of abstract thinking and, consequently, his thinking that women are indeed incapable of such thinking is just an honest mistake since all the evidence he would have been able to gather pointed in that direction. However, if he were to one day come across some counterevidence, we would expect him to adjust his belief about women. If that doesn't happen, then we can say that there is prejudice at work and that he does something ethically and epistemically bad. In the latter instance, Fricker argues that the boy has a negative identity prejudice, and together with Fricker's conception of stereotypes, we come to the definition of the exact prejudice that we find in systematic testimonial injustice Fricker is interested in, and that is a negative identity-prejudicial stereotype defined as follows: "A widely held disparaging association between a social group and one or more attributes, where this association embodies a generalization that displays some (typically, epistemically culpable) resistance to counter-evidence owing to an ethically bad affective investment" (Fricker 2007, 35). As we have seen from the earlier section on stereotypes, stereotypes are something that we use daily and that helps us make sense of the world. As such, we also use them in testimonial exchanges, and the kind of prejudice described by Fricker "distorts the hearer's perception of the speaker" (Fricker 2007, 36). As we also know from the previous section, and as Fricker (2007) writes, stereotypes can affect us even if are not fully aware of it, i.e., they can affect our judgment subconsciously. She illustrates this with an example of a feminist who experiences the influence of stereotypes about women and, consequently, does not take the word of female politicians as seriously as she otherwise would have. This, she claims, supports the idea that testimonial injustice happens all the time.

Consequently, "when prejudicial stereotypes distort credibility judgments: knowledge that would be passed on to a hearer is not received" (Fricker 2007, 43), thus constituting epistemic harm. As Fricker notes, the primary harm of testimonial injustice is that speakers are "degraded *qua* knower, and they are symbolically degraded *qua* human" (Fricker 2007, 44). The secondary harm concerns the practical and the epistemic dimension. An example of the practical dimension would be, if the testimonial injustice were a one-time occurrence, that, for example, a person has to pay a fine if the testimonial injustice was done in a courtroom setting. Or, if testimonial injustice is systematic, it may mean that one wouldn't be able to advance in her career because she would not be deemed competent (Fricker 2007). Fricker explains that the epistemic harm in question, in the cases of systematic testimonial injustice, may concern the loss of confidence in intellectual abilities, which can lead to a decrease in educational performance (let us also remember the issue of stereotype threat described earlier). Or, as she puts it:

> The recipient of a one-off testimonial injustice may lose confidence in his belief, or in his judgment for it, so that he ceases to satisfy the conditions for knowledge; or, alternatively, someone with a background experience of persistent testimonial injustice may lose confidence in her general intellectual abilities to such an extent that she is genuinely hindered in her educational or other intellectual development. (Fricker 2007, 47-48)

Fricker (2007) further illustrates this point by providing an example by Linda Martin Alcoff (2000) of a young female professor who experienced self-doubt because of complaints from a white male assistant until the same happened to her white male colleague. Only then was it concluded that the assistant had problems with authority.

Fricker also thinks that if testimonial injustice is systematic, it can even have an effect on our very identity. As she notes, prejudice can have a kind of a self-fulfilling prophecy (something that we have also mentioned in the previous part of the text, and which was confirmed by empirical evidence from social psychology). Fricker (2007) concludes that "where it is not only persistent but also systematic, testimonial injustice presents a face of oppression" (Fricker 2007, 8).

Now that I have provided a needed background, I will move on to the case of pejoratives, or more specifically slurs.

CHAPTER III: SLURS

3.1. Introduction

In the previous Chapter, I presented the needed background on how stereotypes and prejudice may affect us, as well as presented an overview of Fricker's work on testimonial injustice. In this Chapter, the focus is turned to slurs.

Hate speech, as we have seen in the previous chapter, can take many forms, such as cross burning, which is a symbol of racial hatred. However, one of the most used ways to express hate speech is to use words. It is no surprise that words can have a resounding effect on us. Words can wound, or, as Emily Dickinson has poetically put it, they can hurt us so deeply we may even compare them to physical injuries. But what is so hurtful about words? This is where pejoratives come into the picture, or more specifically, slurs. Slurs are a part of the pejorative cluster; they are a subclass of pejoratives. By analyzing some examples of slurs, I will try to show why we correctly perceive words as being hurtful, and I will portray how they are described in dictionary entries, which will help paint a picture of what lies beneath these terms and what makes them so interesting to philosophers of language. Slurs have been spiking an interest of philosophers of language for decades and there are numerous theories of pejoratives and slurs. Most of these theories fall into two camps: the ones which claim that the slur's content is semantically encoded in the meaning, and the ones which claim that the meaning is pragmatically conveyed. In addition to this, there are some theories that reject this view and that take the non-content based approach (for example, Anderson and Lepore 2013). I will present the most prominent theories of pejoratives to explain how the semantics and pragmatics of pejoratives have so far been analyzed. However, my aim here will not be to introduce a new theory of pejoratives (for now) but to add to the existing theories of pejoratives, namely the ones that focus on the stereotypes connected to one category of pejoratives—slurs.

Even though I will not advocate any particular theory of slurs, it seems to me that there is one crucial aspect of slurs, namely, that when uttering a slur, the speaker is evoking a stereotype. Slurs possessing and evoking stereotypes is not a new idea as it has been put forward by various authors, such as Williamson (2009), Hom (2008), Camp (2013), Mišćević (2016), and others. My view is on the same track, but with a modification. I claim that when uttering a slur, the speaker evokes a prejudice, albeit a specific one—the identity prejudice described by Fricker. This is a novelty I feel best explains how slurs work. However, even though some authors agree there is a stereotype to be evoked in a given slur, the issue lies in the placement of the stereotype, i.e., whether the stereotype evoked in a slur resides in its semantics or pragmatics. Having said that, I will try to offer some arguments (or rather reply

to counterarguments) as to why I am more favorable to the view that a stereotype, or in this case, identity prejudice, could be semantically encoded (even though, as explained, the feature of identity prejudice can work both for semantic and pragmatic theories).

Another important issue that will arise at this point is the issue I briefly mentioned in the Introduction of this thesis. Namely, when presenting examples of slurs, my focus, although not exclusively, will be on gendered slurs for women. The reason for this is twofold: first, it is more practical since I am a woman and I feel more comfortable writing about slurs for women. Second, gendered slurs are in abundance, they are often used, and their semantics and pragmatics are very interesting, as is their historical development. Furthermore, these slurs can be found in every language of the world. The point I will get to later in this thesis is that language reflects our culture, and gendered slurs can serve as an excellent example of that. In fact, as Yule notes, following a long and illustrious tradition, "in the study of the world's cultures, it has become clear that different groups not only have different languages, they have different world views which are reflected in their languages" (Yule 1996, 46). Slurs used to refer to women can tell us a lot about how women are perceived in a given culture. However, gendered slurs have been a subject of controversy. For example, Nunberg (2018) claims that there are no slurs that target women as a group, and Ashwell (2016) argues that gendered slurs lack a neutral counterpart. I will challenge both of these assumptions. In fact, my claim that gendered slurs say something about all women will be one of the central points later in the thesis where I will argue that these kinds of slurs, i.e., group slurs, produce what I will call derogatory-labeling injustice.

In addition, there is one more augmentation of existing theories of slurs I will offer. Namely, in his theory, Miščević (2016) presented levels or layers of slurs (he uses the terms interchangeably). According to him, there would be five levels of slurs: causal-historical, minimal descriptive, negative descriptive evaluative, prescriptive, and expressive. I am sympathetic to this layered view of slurs; however, I will suggest an augmentation of the view—I will include an identity prejudice layer. This serves as an explanation of what Miščević calls the negative descriptive evaluative layer. Namely, the negative evaluative judgment is grounded in negative identity prejudice. Moreover, this layered view of slurs helps us understand the appropriation process, which is something I will refer to in more detail in the fifth Chapter.

3.2. Pejoratives[19]

Pejoratives, in simple terms, as Christopher Hom put it, "express the derogatory attitudes of their speakers" (Hom 2010, 164). Pejoratives can be placed in different clusters. For example, Hom talks about how "the paradigm examples of pejorative words are swear words (e.g. 'damn', 'shit', 'fuck'), insults (e.g. 'dummy', 'jerk', 'bastard'), and slurs (e.g. 'bitch', 'faggot', 'nigger')" (Hom 2010, 164). On the other hand, Kent Bach, at a discussion in Dubrovnik (2014) suggested that we can divide pejoratives into three groups: generic ones, which are associated with a specific genus with examples being *Ni**er* or *Hun*; sub-generic, where a reference to a group is involved with an example of *sheister*; and non-specific pejoratives, such as *asshole* (Miščević 2016). Bach (2018) also distinguishes between group slurs (such as *kike*) and personal slurs, where personal slurs could fit into what Hom calls insults. In my work, I will focus on slurs, or as Bach calls them, group slurs.

We can say that every pejorative has something that can be called perlocutionary[20] potential and perlocutionary range. Perlocutionary potential would be the potential that a certain pejorative has to degrade, offend, etc. The perlocutionary range of a pejorative signifies that a pejorative can be used in various contexts; for example, one pejorative can be used for degrading but also for sarcasm (Austin 1962; Perhat 2012).

In this thesis, I will not be dealing with and analyzing all the pejorative clusters I have previously mentioned. Nor will I be proposing a novel theory of slurs. Here, I will be focusing on slurs since slurs have been found to have a profound effect on their targets, listeners, and the whole of society. Jeshion nicely summarizes the purpose of slurs: "to signal that their targets are unworthy of equal standing or full respect as persons, that they are inferior as persons" (Jeshion 2013a, 232). My aim in this Chapter will be to augment existing theories, specifically those focusing on the stereotypes associated with slurs.

When we think about hate speech, a lot of slurs come to our mind, and I think it would be beneficial to explore them further and analyze them since they can give us an insight into what effect hate speech can have. Slurs have, in recent years, spiked an interest in the philosophy of language, and with good reason. The authors have tried to explain their semantic and pragmatic elements to try to understand how they work. Thus, there are many theories of pejoratives, and in this thesis, I will present just some that are more known and that I find interesting. Before presenting some of these theories, I will first analyze a couple of slurs with the help of dictionaries to get a sense of what a slur is and how it is defined in

[19] Some of the ideas and references I will present in this Chapter are borrowed from and rely on my previous work, namely my MA thesis, 2012.

[20] The perlocutionary act is a speech act that signifies an effect an utterance has on the hearer, for example, frightening or persuading someone.

everyday language. For the analysis, I will concentrate on gendered slurs, specifically, slurs meant to refer to women.[21] But, before moving on, I will need to address issues that were raised concerning gendered slurs since this will make one of the central points later in the thesis.

3.2.1. Gendered slurs

I have pointed out that I differentiate between slurs and insults (based on Hom's (2010), as well as Bach's (2018) distinctions). As it turns out, these distinctions are not always so clear-cut. Usually, it is presumed that slurs target groups even when they are directed at an individual. As Stojnić and Lepore (*forthcoming*) note: "As a result, utterances of slurs—even when applied to individuals— denigrate an entire group; this need not be so for insults and dysphemisms. As Jeshion (2013a) points out, however, this distinction is murky at best" (Stojnić and Lepore *forthcoming,* 6). This issue particularly applies to gendered slurs. Therefore, I would like to take a moment to address an issue that was put forth by Nunberg (2018) where he claims that, in Standard English, there are no slurs for women in general, i.e., that there are no slurs that target women as a group (such as there are for black people, for example), and that slurs for women target only individuals. Perhaps the confusion lies in determining the neutral counterpart of these terms. This idea has been put forth by Lauren Ashwell (2016). By examining gendered slurs for women, Ashwell (2016) has made a point that gendered slurs for women derogate in the same way racial slurs do, but that they are, at the same time, starkly different from other slurs in that it seems they lack a neutral counterpart which is usually a part of the definition of slurs. She concludes that slurs need not have neutral counterparts since keeping this definition would render accounting for some terms nearly impossible. This understanding potentially puts both semantic and pragmatic accounts of slurs at an impasse since both theories require neutral counterparts (Anderson and Barnes 2022). Since my primary interest in this thesis is gendered slurs for women, I will try to offer a potential answer as to why gendered slurs act the way they do. Slurs for women indeed seem to be more complicated to grasp than other, more straightforward slurs for other groups. To understand slurs that target women, we need to look deep into the structure of our patriarchal society. Justina Diaz Legaspe (2018) has offered what seems to be a satisfactory answer to both Nunberg and Ashwell, and I will accept and build on her response. Her goal is to still conceive gendered slurs as slurs and to

[21] Even though my focus will primarily be on gendered slurs, other slurs, such as racial or nationalistic, are also utilized when needed for better elaboration.

keep a connection between slurs and neutral counterparts. Legaspe (2018) explains Ashwell's (2016) point as follows:

> Ashwell notes that there are no referential class-terms both offensively and normatively neutral that single out exactly the same class of people that is derogated by gendered slurs. Take 'slut' for instance: the intuitive candidate for NCSLUT, 'women', is problematic from the start, since not all members of {women} are adequate candidates for being called 'sluts': among others, nuns, wives with impecable behaviour and toddlers should be excluded. Other attempts to find an appropriate neutral counterpart for 'slut' are equally disappointing: 'women who behave in a sexually dissolute manner', 'women who are inclined to behave in a sexually dissolute manner' and 'promiscuous women', they all violate one or both requisites. According to Ashwell, this phenomenon is not restricted to gendered slurs, for it can be observed too in utterances that involve the use of paradigmatic demographic slurs, like Chris Rock's quip (3), in which a clear divide is traced between the reference of the associated neutral counterpart and the reference of the slur:
>
> (3) I love black people, but I hate niggers.
>
> Both cases seem to point out to a failure in ANCT: the thesis can be applied in its negative articulation (only women get to be correctly called 'sluts' and only African-descendants can be called the N-word), but not in its positive articulation (it is not true that all women can be correctly called 'slut' and, according to (3), it is not true that all African-descendants can be correctly called the N-word). (Legaspe 2018, 9-10)

Legaspe builds and adapts what Ashwell has stated, but has not developed further, namely that some slurs impose norms on how certain groups should act. Legaspe's (2018) point is also sufficient to show why gendered slurs are, in fact, still slurs:

> …gendered pejoratives exhibit all the features of slurs presented above: they are directed at individuals in virtue of their membership to a class, and members of it can be offended by their use even if they are not the targeted individual. The differences between gendered pejoratives and non-gendered, demographic slurs are not enough to ban the formers from the set of slurs, since both types of expressions discriminate a whole class of people in virtue of a feature that in itself ought not to mark people as worthy of contempt. (Legaspe 2018, 11-12)

But, the intuition that slurs such as the N-word and gendered slurs such as *whore* are different in some respect still holds. Legaspe offers the following explanation:

> …usage of gendered—and similar—slurs whose reference seems to be always a subset of a neutral class plays, in most cases, a *normalizing* role: derogation via this type of slurs works by pointing to a particular behaviour, behavioural pattern or apparent disposition to behave in a certain way that deviates from what is socially expected from members of the neutral class, with the dual intention of shaming and socially sanctioning the target. (Legaspe 2018, 13)

Legaspe (2018) claims that gendered slurs target a subclass of a certain group. Namely, they target only the members of the target groups that exhibit a behavior that is not deemed acceptable according to the established norm, a behavior she refers to as a P-behavior. She notices that: "a male of any sexual preference can have as many sexual partners as he wants, but a heterosexual female cannot, at risk of being called a 'slut'" (Legaspe 2018, 14). She further explains how: "P-behaviour, by itself, is a *neutral* property not justifying discrimination or social sanction, but when the individual exhibiting it belongs to a certain gender, the community will impose a sanction" (Legaspe 2018, 14). To explain the reasoning behind this, she employs Haslanger's (2012) point on social construction of gender:

> …the social construction of gender is a complex process that launches from gendered bodies and yields a normative interpretation that assigns them with physical features: having female genitals is associated, for example, to lacking physical strength and being capable of giving birth. This interpretation of the gendered body, in turn, gives rise to a social expectance of what these bodies *can* and *cannot* do: it is assumed, for example, that women cannot partake in works that require lifting weight, and women are also expected to become mothers. A particular set of norms ensues that governs gendered behaviour in different dimensions: work, family, social life. These sets of norms are internalized and reinforced by every member of society, including those in the gender class, by means of narrative and sanction of deviation. Therefore, some behaviours that are not even noticeable in members of other genders are singled out as deviations and are negatively marked as worthy of contempt. The P-behaviour thus becomes a deviation from the expected pattern of conduct imposed (and internalized) on members of a certain gender. Because P-behaviour is not even singled out in the behaviour of members of other genders and it is negatively laden for members of the targeted gender, it is difficult to characterize it in a neutral, purely descriptive way: the P-behaviour associated to being called 'slut' can only be articulated in a way that conveys social sanction directed at women. (Legaspe 2018, 14-15)

Finally, Legaspe (2018) makes a point that indeed gendered slurs do have neutral counterparts and will build on this further. Namely, Legaspe (2018) claims that every member of a group labeled with a gendered slur can *potentially* exhibit P-behavior. This is also why every member of a group has a right to protest the use of a gendered slur when it is directed at an individual:

> But in real life even women with the most impeccable behaviour should protest when some other woman is called a 'slut'. This reaction—when one is not the intended target of the normalizing slur—is the result of the combination of linguistic competence and a deep understanding of gender as a social construct. Together, they determine a reception of the slur such that the speaker knows that whenever a woman is called 'slut', a normative system is being enforced—and reinforced—on all women, forbidding them to P-behave. (Legaspe 2018, 16-17)

This is an important point that can be reinforced by utilizing Fricker's notion of identity prejudice. Although I will explain Fricker's points more in-depth in the next section, I will use her notion of identity prejudice at this point because it can provide a fuller understanding of why gendered slurs indeed do refer to all women.

Identity prejudice is related to our social identity (our social identities are social conceptions we share in the collective social imagination that govern what it means to be a man, a woman, etc.), and they function as tracker prejudices that follow us through every social dimension (Fricker, 2007). As she explains, the so-called negative identity-prejudicial stereotype is: "A widely held disparaging association between a social group and one or more attributes, where this association embodies a generalization that displays some (typically, epistemically culpable) resistance to counterevidence owing to an ethically bad affective investment" (Fricker 2007, 35). Even though there are positive identity prejudices such as that women are intuitive (Fricker, 2007),[22] Fricker, as will I, focuses on negative identity prejudice, such as that black people are prone to violence. Through socialization, we acquire social norms and customs, we take up our assigned gender roles and we also encounter stereotypes and prejudice. Some of these stereotypes, as many authors and researchers have shown, may linger in our minds even when we are not fully aware of them.[23] As already stated, to fully understand gendered slurs for women, we need to delve deep into the structure of our society. Women in our society are perceived as sexual beings; it is almost as if being perceived in that way is part of every woman's identity. Granted, one may immediately think of cases where this, *on the face of it*, isn't so. For example, little girls or old women, our grandmothers, and so on. However, as I emphasized, this appears to be so *only* at first glance.

[22] Although this could also be a negative identity prejudice at times.

[23] As shown in Chapter II.

Once we take a deeper look and dissect the underlying assumptions of this, we will discover that indeed sexuality lurks behind this and is tied to women's identities even in such cases. These are the cases where women are stripped of their sexuality completely. In these cases, when they are not perceived as women who can be sexual objects, they are considered to be almost useless and obsolete, i.e., it is as if they are not considered to be women at all. Margaret Atwood's dystopian novel The Handmaid's Tale (and later the series) has a nice depiction of this: Atwood introduced the term "Unwomen" to refer to women who have no use in society because they are infertile. That would, in the real world, for example, be the case of much older women. Even the slurring term *hag* refers to a woman considered to be old and unattractive.[24] Another real-life, personal example of this would be when I have participated, on several occasions, in conversations about friendships where my male acquaintances would claim that they can be very close friends with their female friends because they don't perceive them as women. Furthermore, being perceived as a little girl is often the opposite of being perceived as a woman where *little girl* is synonymous with being innocent and, of course, not being perceived in a sexual way. The perception of becoming a woman is tied to sexuality: you "become" a woman once your menstrual cycle begins or once society begins to perceive you in a sexual way. These are just some of the examples of how our society perceives women through their sexuality, and granted, this is not often so clear-cut. Nonetheless, the punchline that strengthens Legaspe's (2018) view is this: *women should not P-behave because of identity prejudice*. So, the underlying assumption is the negative identity prejudice: women should be ashamed of their sexuality, and exhibiting promiscuous behavior is not lady-like—which is something a woman should aspire to be, exhibiting some kind of sexual behavior devalues a woman's worth in society. And, as Legaspe (2018) correctly notices, all women have the potential to exhibit P-behavior *because* identity prejudice applies to all women. Sexuality has long been perceived as something that all women should be ashamed of, and so there are a lot of slurs that target just that. Slurs for women, in most cases, target their sexuality. Therefore, I argue that slurs for women do target women as a group—namely, they target women's sexuality since being perceived as a sexual agent is part of being a woman; it is a part of every woman's social identity.[25] Indeed, we can target just an individual if we want to say something about her character, but by doing so, we are also implying something about the group she belongs to. Namely, we are implying something about other women in general. So, even if a slur is directed at an individual target (for example, a woman somebody refers to as a *whore*), in addition to making an evaluative descriptive judgment about this particular target, the speaker is also saying something about

[24] Taken from: https://dictionary.cambridge.org/dictionary/english/hag

[25] As Fricker explains, the conceptions of social identity are "conceptions alive in the collective social imagination that govern, for instance, what it is or means to be a woman or a man, or what it means to be gay or straight, young or old, and so on" (Fricker 2007, 14).

the group the individual target belongs to and thus indirectly gives an evaluative judgment about the group the target belongs to, as well. For example, in case the speaker refers to an individual woman as a *whore*, they are also saying something about women in general; namely that P-behaving is forbidden for all women because of negative identity prejudice that women should be ashamed of their sexuality. If being perceived as a sexual agent is tied to being a woman, that means that any slight manifestation of her sexuality can be deemed as promiscuous behavior, which is, of course, viewed negatively because women should be ashamed of their sexuality. In that sense, I understand the term *promiscuous* very broadly. Promiscuous behavior can, for someone, be when a woman is dressed revealingly, or when she looks at someone in a certain way, or it can be the way she walks, talks, or just the way she *is*. Since women are viewed as sexual agents and since this is tied to their social identities, any number of behaviors women exhibit can be deemed as promiscuous. And, as Legaspe (2018) argues, P-behavior (in this case promiscuity[26]) is neutral, and only becomes deviant behavior once it is applied to certain groups.

As Nunberg (2018), Ashwell (2016), and Legaspe (2018) have noticed, there indeed is something different in gendered slurs as opposed to slurs such as the N-word, which Legaspe (2018) calls demographic slurs. As already mentioned by Legaspe (2018), gendered slurs are tied to P-behaving and what she calls demographic slurs, such as the N-word, seem to have an underlying universal identity prejudice tied to an ascribed characteristic (being black). Gendered slurs refer to P-behaving, and P-behaving is tied to negative identity prejudice internal to being a woman.

Clearing the air regarding these issues was important because later in the text, I will argue that each type of slur produces what I call derogatory-labeling injustice, even though one type of slur (the N-word) may be regarded as hate speech, while another (gendered slurs) may not. However, there is a need to address both types of slurs since both can produce various harms I will discuss in Chapter IV.

To reiterate, the view on women's sexuality, as already stated, is a very complex one, and many feminist authors, such as Robin Lakoff (1973), feel that "the marginality and powerlessness of women is reflected in both the ways women are expected to speak, and the ways in which women are spoken of" (Lakoff 1973, 45). Gendered slurs for women can serve as a prime example of Lakoff's claim. Even though women's sexuality is not the topic of this thesis, by analyzing some of the slurs, it is easy to notice that many slurs that refer to women refer to their promiscuity. Also, as Hughes notes, there is an obvious imbalance between the number of terms applied to women's promiscuity as opposed to men's. Furthermore, Hughes

[26] Some may claim that promiscuity is never neutral. However, as I explained, I view promiscuity as a very broad term, and in that sense, it *should* be regarded as neutral because it is a subjective matter of what one deems to be promiscuous behavior.

also notes that "there is the semantic fact that unfavorable terms for women outnumber positive terms by a proportion of about five to one" (Hughes 1991, 225). Also, it seems there is a so-called *angel/whore* dichotomy when it comes to terms that refer to women where the terms of praise for women are *angel, virgin, maiden, or goddess* while derogatory terms are *whore, witch, bitch,* etc. (Hughes 2006).

Hughes (2006) offers an exhaustive list of slurs that exist and have existed in past to describe women's promiscuity, such as: "*whore* and the obsolete *quean*, from Anglo-Saxon times, as well as *harlot, strumpet, concubine, call girl, hooker, tart, tramp, moll, hustler, streetwalker, pickup, scarlet woman, fallen woman, woman of the streets, woman of easy virtue, lady of the night,* and *escort"* as well as "*fast woman, hussy, doll, inamorata, siren, gypsy, minx, vamp, wench, trollop, coquette, bint, crumpet, floozy, scrubber, slag, groupie, nympho,* and *slut*" (Hughes 2006, 363). He also mentions some other archaic terms, which I will not include here. Of course, I will analyze some of the most used terms from the list.

3.2.2. Analysis of selected slurs

Let me start with a gendered slur *slut*. According to Oxford Advanced Learner's Dictionary, the word slut has two meanings: 1) a woman who has many sexual partners and 2) a woman who is very untidy or lazy. The first meaning has a strong sexual reference directly linked to promiscuity. But, in the second meaning, this sexual reference is lost even though it still has a negative meaning. As we can see, both uses, which are negative, apply to women.

Let me continue with another often-used term, *whore*.

Oxford Advanced Learner's Dictionary lists two meanings: 1) (old fashioned) a female prostitute, and 2) (taboo) an offensive word used to refer to a woman who has sex with a lot of men. The first meaning is literal, where the word refers to a female who sells her body for money. In the second meaning, the reference to money is lost, and the word just refers to a promiscuous woman, so we can say it acts as a half-dead metaphor.

As Hughes (2006) notes, the word *whore* has an interesting etymology:

> Although found in Anglo-Saxon, whore is recorded late in comparison with the related Germanic languages. The etymology is fascinating, since whore has cognate forms in Latin carus, "dear," and Old Irish cara, "a friend." It first appears in the form hore, subsequently huir, indicating the pronunciation "hoor" or "hooer," which continued into the nineteenth century, and as the OED noted, "may be adopted . . . when we

> wish to soften the effect of a coarse word." The spelling with wh- became current in the sixteenth century. (Hughes 2006, 493)

The word is interesting because it has not, throughout centuries of use, lost its power. Recently, the third meaning has arisen where the word applies to "anyone who sells out their principles" (Huges 2006, 493). In this meaning, the word no longer strictly refers to only women but can refer to both sexes, and the sexual component present in the first two meanings is now lost, while the negative component is still there. Here, we can also mention that the Urban Dictionary offers yet another most recent use of the word where it refers to somebody or something that is excessive and repetitive, which is very annoying. An example would be *photo-whoring*, which refers to excessively taking photos (most likely of oneself) that other people find annoying. In this expression, the link to the literal meaning of prostitute is obviously lost, as is the sexual reference, although we can say that some of the link to promiscuous behavior is still here because promiscuous behavior is connected to excess and repetitiveness.[27]

I will finish this short analysis with a term that today does not necessarily relate to promiscuity, even though the link to it can still be found. The word *bitch*, as Hughes (2006) notes, was indeed used in the past to refer to promiscuous women as an extension of a female dog in heat. The word has an interesting history and development, or as Hughes explains, the word "has the longest history among animal terms as an insult, extending from the fourteenth century to the present, during which time it has steadily lost force through generalization" (Hughes 2006, 23), and, I would add, gained new meanings along the way.

As Oxford Advanced Learner's Dictionary lists, there are several meanings of the word: 1) the first one is literal, where the word refers to a female dog, 2) the second meaning is an offensive way of referring to women, especially an unpleasant one, 3) in the third meaning the word refers to a thing that causes problems or difficulties (as in: "Life is a bitch."), 4) finally, the fourth meaning listed in the dictionary is when one complains about somebody or something (as in: "We've been having a bitch about our boss."). There are also a lot of derivative forms of the word, such as the adjective *bitchy*, which the Oxford Advanced Learner's Dictionary explains as saying mean things about other people (as in "bitchy remarks"). The dictionary also lists the newer derivative meaning of the word where the meaning transformed into a positive one where the adjective *bitching* means something very good. As we have seen, the word itself started with the literal meaning, which then, as Hughes noted, referred to a female dog and was later used to refer to promiscuous women. Even though the link to promiscuity is not evident in these new meanings, we can still find traces of that use today, for example, in pornography or in expressions like "make somebody one's bitch" in the sense of making somebody submissive. Nevertheless, in almost all meanings,

[27] The analysis of the word *whore* has also been used in my MA thesis (Perhat 2012).

there is still a negative connotation. As we have seen, only the second meaning listed in the dictionary refers to women, while other meanings (except the literal one) can refer to both sexes and situations with a negative meaning.[28] To conclude, *bitch* is a term that is halfway on becoming reclaimed, just as the term *queer* is today and therefore is not a paradigmatic example of the full force of gendered slurs.

Now that I have analyzed three pejorative words, or more precisely slurs, to better understand how slurs are defined in dictionaries, I will turn to various theories of pejoratives that try to explain their semantics and pragmatics. I will not be dealing with all of the theories because, in recent years, there has been an increasing interest in pejoratives in the field of the philosophy of language, meaning that there are too many theories for this thesis to take into account. I will focus on the ones that gained the most interest and the ones I find more plausible than others.

[28] The analysis of the word *bitch* has also been used in my MA thesis (Perhat 2012).

3.3. Theories of pejoratives

As emphasized, there is an abundance of theories of pejoratives, but for the purpose of this thesis, I will simplify the division and divide them into two camps, following the footsteps of Miščević's (2016) and Hom's taxonomies (2010). Since Miščević (2016) has done great work in systematizing theories of pejoratives, I will be relying mostly on his work. When considering various theories of pejoratives, one must keep in mind that all of the theories have their strengths and drawbacks. Pejoratives form a complex phenomenon that is not easy to unpack; when discussing pejoratives (or slurs, since slurs are what primarily interests us in this thesis) one has to account for their semantics, pragmatics, socio-linguistic and political effect they may have. Nevertheless, I believe that, when taking all these features that pejoratives (slurs) carry, some theories work better than others. This is why I will not be dealing with particular theories in detail (nor with the criticism presented for each theory) but portray only the main ideas of how some of the most prominent theories think of pejorative words, which can help us understand how pejoratives work in real life, and when used in hate speech.

In the broadest sense, we can divide the theories into two camps: non-semanticist (or expressivist) and semanticist theories of pejoratives.

3.3.1. Non-semanticist theories of pejoratives

There are several non-semanticist theories of pejoratives, and in the broadest sense, we can say that non-semanticist theories claim that pejorative content cannot be reduced to semantic content (Hom 2010).

To understand where the non-semanticist theories gained their ideas, I will briefly borrow from Miščević's (2016) analysis of Frege. Namely, in "On Sense and Reference", Frege claimed that the difference between a pejorative and its neutral counterpart is in the tone. By giving an example of a *dog* and a *cur*, he explained that the tone of the two is different. In the former, the tone is neutral, but in the latter, the tone is a negative one, where we express a negative attitude. Furthermore, the tone, according to Frege, does not participate in determining the truth value of a preposition (Miščević 2016). As Miščević (2016) also explains, this is where the non-semanticist, or the expressivist theories as Miščević calls them, borrow their ideas.

One of the examples of such theories would be expressivism and gesturalism.

Robin Jeshion, for example, supports a version of an expressivist view. Or, as she explains:

> My view shares a key feature with extant expressivist analyses of slurs' offensiveness: one component of slurs' semantics involves the expression of contempt toward targets on account of membership in the socially relevant group. Expressivist semantics of slurring terms are tailor-made to capture slurs' common and conventional capacity to derogate. For according to such theories, on an occasion of use of a slur with its literal meaning, a competent speaker semantically expresses contempt toward the target, either an individual or a group (or both), on account of being a member of a certain socially significant group. (Jeshion 2013a, 308-309)

Jeshion (2013a) maintains that slurs express contempt towards group members because they are in that group or, in other words, they express the speaker's attitude. She finds similarities in semantic properties between slurring terms and expressives, such as intensifiers, exclamatives, or performative expressives. So, slurs would have similar semantic properties to words such as "damn", "ouch", or "wow", where we express our emotional attitude. Besides the expressive component of slurs, Jeshion mentions one more, the identifying component, and expounds how "the expressive and identifying components explain slurs' common and conventional capacity to derogate. As a matter of their semantics, slurs function to express the speaker's contempt for the target in virtue of the target's group-membership and that his target ought to be treated with contempt in virtue of that group-membership, because what the target is, as a person, is something lesser, something unworthy of equal or full respect or consideration" (Jeshion 2013a, 319).

Another example of a non-semanticist view is that of Jennifer Hornsby. In her article "Meaning and Uselessness" (2001), she argues that derogatory words are useless in that we, if we are not racists, bigots, and so on, simply have no use for them. She further explains that we can think of slurs in the following way:

> It is as if someone who used, say, the word 'nigger' had made a particular gesture while uttering the word's neutral counterpart. An aspect of the word's meaning is to be thought of as if it were communicated by means of this (posited) gesture. The gesture is made, ineludibly, in the course of speaking, and is thus to be explicated, as the socially significant thing it is, in illocutionary terms. The gesture has no life of its own, independently of the use of the derogatory word, so that there is nowhere else to look, to appreciate its significance, than to uses of the word (*pace* Hare). (Hornsby 2001, 140-141)

We can consider her theory as gesturalism; the slur would, in simple terms, consist of a neutral counterpart plus an added gesture of contempt.

Here, I have presented two theories that can be considered as non-semanticist theories because of the view that what constitutes the semantics of a pejorative is minimal. There has been a lot of criticism of said theories, mainly, of course, from the semanticist camp. The negative part we associate with slurs would not be a part of their semantics but rather a part of their pragmatics.

According to Hom (2010), there are a couple of features of pejoratives that every pejorative theory needs to take into account, namely: expressive force, force variation, taboo, historical variability, syntactic variability, generality, ineffability, the deduction puzzle, the balanced construction constraint, the infixation constraints, the content dichotomy puzzle.[29] Hom argues that some non-semanticist theories of pejoratives fail to account for some of these features. The main problem Hom sees with non-semanticist theories is their inability to deal with the embedding of pejoratives in negative sentences or questions. In such cases, the speaker is not expressing her negative attitude, and according to some non-semanticist accounts of pejoratives, namely expressivist theories, such negative attitudes should accompany pejorative terms. Furthermore, sentences that use slurs and sentences that use their neutral counterpart would, under expressivism, say the same thing because, according to expressivism, what is expressed is not part of the meaning of a slur. Miščević (2016), on the other hand, while criticizing expressivist theories, stresses the cognitive work that is involved while using and deciphering slurs, which he thinks is a good indicator of semantic structure.

3.3.2. Non-content based account[30]

Before moving to semanticist theories of pejoratives, I would like to turn my focus on one more interesting theory that ascribes to semantic minimalism, and that is the theory presented by Anderson and Lepore. Nenad Miščević and I have been working on a reply to their theory written below which we also addressed in our book A Word Which Bears a Sword (2016).

Lepore and Anderson, as they explain it "defend a non-content based view. According to us, slurs are *prohibited* not on account of offensive *content* they manage to get across, but rather because of relevant edicts surrounding their prohibition" (Lepore and Anderson 2011,

[29] For further elaboration on this see Hom (2010).

[30] Objections raised here were presented in Miščević, Perhat (2016).

17). They claim that no content-based account of pejoratives can be correct and they propose that slurs are prohibited words and that "offensive words are generated by word-taboos" (Anderson and Lepore 2011, 17). So what they say is that "no matter what its history, no matter what it means or communicates, no matter who introduces it, regardless of its past associations, *once relevant individual declares a word a slur, it becomes one*" (Anderson and Lepore 2011, 16), and relevant individuals being members of targeted groups, but needn't be. Lepore and Anderson also stress how "the dominant group's use of the expression might be a vivid reminder of the relation of oppression in which the subordinate group is situated" (Lepore and Anderson, 2013). But isn't that, the relation to oppression, the part of the offensive content the slur carries in the first place? Or, as Miščević and I (2016) have pointed out through an example:

Group B is oppressed by group A which uses the term T for B's. It can signal oppression in two ways:

1.) the use of T is itself part of oppression and thus T is demeaning.

2.) the use of T is neutral and there is only historical association linking T with otherwise distinct fact of oppression.

So, let us now see which way seems to be more plausible, 1.) or 2.):

1.) It seems extremely implausible that A would use a neutral term for members of the group they severely oppress; the non-neutral use makes the reaction of B's to the term normal and rational.

2.) If T is neutral, then the reaction of B's is contingent and non-rational, which makes 2.) extremely implausible.

It is obvious that 1.) is incompatible with Lepore and Anderson's theory (Miščević and Perhat 2016, 132).

The infamous example is the word *Ni**er*. Huges (2006) writes that this word derives from the practice of slavery and he himself explains: "This primal link with slavery is obviously vital, since it embodies in an intensified fashion the demeaning roles of servitude and of being an outsider that have characterized the early roles of black people in Western society" (Hughes 2006, 327). It is also a known fact that this word was widely used in the past and that it wasn't prohibited at all even though it was demeaning (Miščević and Perhat 2016, 132). This point was also emphasized by T. Williamson in a conference in Dubrovnik (2012). In the same conference (2012) Williamson also pointed out that there are examples of mild pejoratives which nobody prohibits and that we wouldn't consider them to be taboo (his example was the word *Pom* which is a mild Australian pejorative for British people).

I have already presented some questionable points that can be attributed to Lepore and Anderson's account of slurs, and here I will shortly present some of the problems expressed for non-semanticist views where I will mainly rely on Hom's and Miščević's views.

3.3.3. Semanticist theories of pejoratives

The semanticist theories of pejoratives claim that the negative content of the pejorative is part of its semantics, which makes pejoratives rich in content. In other words, the negative part of the slur would not just be the expression of contempt but instead would be situated in the very meaning of the term.

In this part, I will present two semanticist theories that I believe are the most probable and which, to my mind, offer plausible explanations of the ways how pejoratives work. These are the theories of Hom, May, and Miščević. I will begin with Hom and May's theory and then move on to Miščević's theory.

3.3.3.1. Whores as unicorns?

Before presenting Hom and May's theory (2018), I will briefly discuss Hom's theory of pejoratives, which he put forward in 2010, calling the theory Thick semantic externalism. Hom explains it as:

> For any slur D, and its neutral counterpart N, the semantic value for D is a complex property of the form: *ought be subject to p*1 + ... + p*n because of being d*1 + ... + d*n all because of being N**, where p*1, ..., p*n are deontic prescriptions derived from the set of racist social practices, d*1, ..., d*n are the negative properties derived from the racist ideology, and N* is the semantic value of N. For example, the slur 'chink' expresses a complex, socially constructed property like: *ought to be subject to higher college admissions standards, and ought to be subject to exclusion from advancement to managerial positions, and ..., because of being slanty-eyed, and devious, and good-at-laundering, and ..., all because of being Chinese*. Basically, to call someone a D is to say that they ought to be subject to discriminatory practices for having negative, stereotypical properties because of being an N. (Hom 2010, 30)

In short, Hom (2010) explains that "thick semantic externalism is the view that pejorative content is a socially determined, truth-conditional prescription" (41) and that they are offensive in every context because, in every context, they carry the negative valence in their semantics (even though, it needn't be the case that in every context somebody will be offended because offensiveness is a psychological reaction which may or may not occur). Let me now move on to a newer theory presented by Hom and May (2018), where they claim that we can treat slurs as fictional terms.

Hom and May (2018) proposed that slurs are fictional terms. They have elaborated on this by providing an allegory of the Middle Ages when people used to believe in the healing effects of unicorn horns. They have built an entire mythology around unicorns as a result. Of course, unicorns do not exist. What people thought to be unicorn horns were, in reality, narwhal tusks. In that sense, as Hom and May explain, the unicorn horn has a null extension because unicorns do not actually exist.

Hom and May (2018) further expand on this by saying that there are two ways we can look at the truth value of fictional propositions. On the one hand, fictional propositions are materially false; for example, there is no such thing as a unicorn. On the other hand, they can be fictionally or mythologically true. When, for example, we discuss a fictional book about unicorns, we would want to say that it is mythologically true that unicorns indeed have horns. But Hom and May press further on the matter by expanding the debate to propaganda, noting that it is dangerous because of the consequences it can produce. In this case, there is a negative myth surrounding a slur, which is perpetuated by speakers using this term and by justifying the negative treatment of groups of people. So, just like there is a developed mythology around narwhal tusks, there is also a negative mythology (ideology) around slurs. Thus, one can believe there are *whores,* just like one can believe there are unicorns, but both terms have null extensions according to Hom and May, even though fictionally they have non-null extensions (Hom and May 2018).

Building on this view, Hom and May developed their Moral and Semantic Innocence theory in which they claim that:

> pejoratives express a semantic component that is represented as PEJ that is a second-level concept that takes first-level group concepts (e.g. being Jewish, being Chinese, being African-American, etc.) as inputs and maps them to first-level concepts (e.g. being a kike, being a nigger, being a chink, etc.). These in turn map to False for every argument. They do so precisely because of the negative normative judgment that the PEJ concept expresses - something like: ought to be the target of negative moral evaluation because of being a member of G, where G is the first order group concept term. (Hom and May 2018, 5)

It follows that since "no-one ought to be negatively evaluated on the basis of their group membership, pejorative terms like 'kike' have empty extensions" (Hom and May 2018, 5). Thus, $whores = \{\!\{\emptyset\}\!\}$.

But what are the instantiations of group G? Hom and May argue that group membership cannot be morally evaluable, just as set membership. However, there are some restrictions as to what kind of groups are a value of G, and this restriction is determined by active ideologies. If sexism and misogyny die out, *whore* or *slut* would not be pejorative terms: "the life of an ideology supervenes on the life of a pejorative term" (Hom and May 2018, 6). Or, as they explain:

> Thus, the answer to the question at hand—What are the criteria for choices of *G* such that there will be a pejorative term with the meaning (sense) PEJ(*G*)?—is that it is reserved for groups that for whatever odious reasons have associated with them an unjust, hateful, or discriminatory ideology that is culturally ingrained within society. Targeting a group in this way creates an illusion, a fiction; pejoratives are terms of these fictions. (Hom and May 2018, 8)

Of course, there are certain questions that arise from such a view. The most important one was already addressed by Hom and May. There are people that should indeed be negatively morally evaluated, for example, Skinheads, Nazis, etc. To further explain:

> In a certain sense, mass murderers form a group, and being a mass murderer justifies negative moral evaluation in virtue of the action one must take in order to become a member of that group. Being a member of that group, however, does not justify being the target of pejoration. *Qua* group, mass murderers are no different than any other group in this regard: without a supporting ideology, there can be no pejoration. (Hom and May 2018, 10)

In considering offensiveness, Hom and May are clear that offensiveness is behavioral, meaning that what is offensive for me may not have the same effect on someone else. Therefore, according to their view, a pejorative "is an expression of moral contempt, of negative normative judgment, not offensiveness" (Hom and May 2018, 10), and offensiveness is not part of the meaning of a slur. They explain:

> A directly pertinent illustration arises by conceiving of a social-historical context whereby an oppressed group has fully internalized the ideology of their oppressors creating a false consciousness of inferiority. No one is offended by pejorative terms used to refer to this oppressed group because everyone believes the surrounding ideology; in short, everyone believes members of the oppressed group are morally inferior, including

> those members themselves. Clearly, we can have pejorative terms in this context where the racist ideology has been so completely internalized that no one takes offense. Because there is no offense in this kind of scenario, there is also no taboo. Even members of the targeted group believe that they are intrinsically inferior and so no one objects to uses of these words. (Hom and May 2018, 12)

But, even though no one might take offense by a slur because offensiveness is behavioral, the negative moral judgment is still rooted in the meaning of a slur, whether someone takes offense by it or not (Hom and May 2018). As far as truth conditions of pejoratives go, Hom and May explain as follows:

> By any account of the truth-conditions for sentences containing pejoratives, a simple sentence like "Max is a kike" is true if and only if Max falls under the concept of being a kike. But since no one falls under the concept of being a kike, from these truth-conditions it follows that "Max is a kike" is false. (Hom and May 2018, 15)

To further explain, the sentence "Jane is a slut." is false because there is no such interpretation where one would fall under the concept of being a slut, i.e., there are no sluts.

There is one more question that comes to mind when thinking about Hom and May's account and that is the question of reference. One might ask if slurs have a null extension, to what do they refer? Hom and May (2018) explain as follows:

> On MSI, the meaning of "kike" is fixed, relative to a negative ideology; it is this fixation of the meaning of PEJ(*G*) for Jews as the instantiation of *G* that we have glossed as the concept *ought to be the target of negative moral evaluation because of being a Jew*. But this fixation contravenes an *a priori* moral principle, and so reference-fixing is immoral. It is with this very immorality that the roots of null extensionality lie. On our view, this immorality is in the same league as the unscientific convention that fixes the reference of "unicorn," and more generally with the overall unreality of fiction. (Hom and May 2018, 28)

To further expand on this, it does seem that one stumbling stone in recent theories of pejoratives is how to secure the reference of a pejorative. For example, we could ask what or who does the term "Ni**er" refer to. It seems that when one utters a pejorative word, one is referring to someone, but the question is, to whom? And, furthermore, what sets the truth value of such sentences? We as hearers would intuitively know that the speaker who uses the pejorative, say "Ni**er", is thinking of a black person, but, if we are not bigots, we would not use the same word and would not agree with the speaker that the target of their speech is, in fact, a Ni**er. If we did, that would make us racists. But, if we do not think the target of

such speech is a Ni**er, then who is the speaker referring to? These are the questions about the slurs' references that many philosophers of language tried to address in order to resolve the puzzle.

Some philosophers of language would claim that pejoratives do refer. They would differ in terms of what secures the reference. For example, Jeshion considers that the reference of pejorative terms amounts to its counterpart, and so she would propagate a sort of minimal reference theory of pejoratives. So, the idea would be that the term "Ni**er" refers to a black person (or, more narrowly, an African American). This is a bit tricky because it would mean that sentences like "Max is a Ni**er." are true because Max indeed is an African American.[31]

Others (Williamson 2009; Miščević 2016) will claim that the negative evaluative part is not part of the reference but also that pejorative terms do refer. For Williamson (2009), the reference is part of the conventional implicature, and for Miščević (2016), the "reference is partly determined by causal chain" (99). For Miščević (2016), the reference is secured thanks to this causal part, but also thanks to the minimal descriptive part.[32]

Other theories, such as the one proposed by Bach (2018), claim that we should refrain from assigning truth value to such sentences since one part of the pejorative sentence, like "Max is a Ni**er.", would be true (that Max is an African American) and the other one (that Max is bad because of this), would not be true. Jeshion criticizes this failure to secure the reference of pejorative terms and sees this as a potential problem for semantic theories.

3.3.3.2. Pejoratives as social kind terms?

Let me now continue with another semantic theory of pejoratives, the one proposed by Nenad Miščević, a theory which he calls The Negative Hybrid Social Kind Terms theory (Miščević 2016). Miščević, going along with the semanticist tradition, ascribes the bad material of pejorative terms to be part of the meaning. Miščević describes his theory as follows:

> ...pejoratives, say "N", are negative (derogatory) social kind terms, with a hybrid nature. Their reference is partly determined by the causal chain: the target group G has been called by somebody "N", the name has been transmitted to the present users, and it refers to the group G and its members. Their descriptive senses have neutral material (given by a neutral

[31] I have already mentioned that Jeshion is fine with the idea that such sentences would be rendered as true.

[32] Fuš (2016) criticizes Miščević's account of reference of pejoratives.

> description: „German", "female", "gay"), and bad material (primitive, hateful, stupid, etc.) plus more; we shall return to the prescriptive and expressive components in a moment. Let me call the proposal the Negative Hybrid Social Kind Terms Hypothesis (NHSKT hypothesis). (Miščević 2016, 99)

To further explain, Miščević provides an example taken from superstitious beliefs about medicine men. Say that one utters the sentence "He is a medicine man."—would that sentence be true or false? Miščević explains that the sentence would be literally false if taken in its superficial meaning since the medicine men do not have magic powers. But, if that sentence were uttered by, say, an anthropologist, then the proposition would be true since the man in question does preform some activities and is believed by other tribe members to have magic powers.

Miščević endorses a view that pejoratives have thick semantic meaning, i.e., that they are very rich in their semantic content, so we can say that he endorses the maximal semanticist proposal. According to his proposal, we could pinpoint five levels of pejorative content: casual historical, minimal descriptive, negative descriptive-evaluative, prescriptive, and expressive. Only the expressive level would be part of the pragmatics of pejoratives, while other levels would be part of the semantic content. Thus, accordingly, we can sketch out these layers as follows, taking the word *whore* as an example:

Table 1

Example: *whore*		
	Level	**Content**
Semantic part	Casual-historical	Someone called them thus
	Minimal descriptive	Woman
	Negative descriptive-evaluative	Bad, sinful, unethical
	Prescriptive	To be avoided, not to get romantically involved
Pragmatic part	Expressive	Yuck!

To decode the above table a bit further: according to Miščević's theory, when somebody utters the word *whore*, what they mean is that they are referring to a woman who

is immoral, sinful, and unethical and should thus be avoided, and the speaker is disgusted by her. As per the sketch, it is evident that Miščević (2016) endorses a plurality view. From this plurality of propositions, according to Miščević, it is the context that determines the relevant proposition. Thus, sometimes, the sentence with a pejorative would be true, and sometimes false, depending on what the focus is on in the said sentences. So, if the focus is the bad material encoded in the pejorative, the sentence would be false. Then, accordingly, if the bad material is not the focus, the sentence could be true. So, for the term *whore*, if we were to take the word's derogatory content alone as the focus point, the word would fail to refer. What secures the reference is both the casual link and the minimal descriptive link.

To conclude, it seems that the two theories I have presented as prime examples of semantic theories of pejoratives have the similar idea that pejoratives are thick concepts with rich semantic content. Both theories consider the negative valence to be part of the meaning of a pejorative, but they differ in terms of reference. For the purpose of this thesis, these differences are less important.

3.4. Unpacking the content: slurs and stereotypes

Before further unpacking the content, I will agree with Jeshion (2013a) that one needs to distinguish between various types and various uses of slurs. First, as Jeshion (2013a), I will also consider only literal uses of slurs in my analysis. That means that when used in a literal sense, slurs "are used to reference the group referenced by their neutral counterpart and as weapons, to derogate that group" (Jeshion 2013a, 315). Slurs used as weapons is an important distinction Jeshion rightly emphasizes. As Jeshion (2013a) explains:

> Slurring terms are used as weapons in those contexts in which they are used to derogate an individual or group of individuals to whom the slur is applied or the socially relevant group that the slur references. The following sentences, as spoken by the racist, anti-Semite, homophobe, etc., are all weapon uses of group-referencing slurring terms:
>
> [1] Yao Ming is a Chink.
>
> [2] Barbara Streisand is a Kike.
>
> [3] He is a faggot.
>
> [4] You Kike!
>
> [5] The actors in that play are all Niggers.
>
> [6] Hire one of the Spics over there.
>
> [7] The movie was about a bunch of Chinks.
>
> These uses are fruitfully contrasted with non-weapon uses, utterances of sentences containing occurrences of (non-appropriated) group slurring terms yet the slurring terms themselves are not being used to derogate or contemn any groups or individuals on the basis of their group membership. (Jeshion 2013a, 313)

These literal, weapon uses of slurs that refer to a group (as distinguished by Bach (2018), Hom (2010), and Jesion (2013a)) will be the focus of my work.

Considering the above distinction, I claim that these most vicious uses of slurs have one crucial aspect, namely, that when uttering a slur, the speaker is evoking a stereotype. Slurs possessing and evoking stereotypes is not a new idea, as it has been put forward by various authors, such as Williamson (2009), Hom (2008), Camp (2013), Jeshion (2013a),[33]

[33] In her 2013a and 2013b articles, Jeshion actually argues against stereotypes being semantically encoded in a given slur, and she argues against the claim that stereotypes will be evoked in every utterance. As I present my case, it will become evident that the negative identity- prejudicial stereotype Fricker describes is different

and Miščević (2016). My view is on the same track, but with a modification. I, too, claim that a stereotype is evoked by uttering a slur, albeit a specific one—the negative identity prejudice described by Fricker. This is a novelty I feel best explains how slurs function, which I will elaborate on later in the text. However, even though some authors agree there is a stereotype to be evoked in a given slur, the issue lies in the placement of the stereotype, i.e., whether the stereotype evoked in a slur resides in its semantics or pragmatics. And, even so, stereotypes being central to slurs has been an issue of a lot of controversy and some compelling counterarguments were presented. I will offer a brief discussion and an overview of the issue of whether a stereotype could be semantically encoded. I do not hold a view that identity prejudice is semantically encoded because identity prejudice can work both for semantic and pragmatic theories. In fact, this distinction is not essential to what I am about to claim later in the thesis—slurs produce derogatory-labeling injustice in both cases: whether the identity prejudice is semantically encoded or whether it is pragmatically conveyed.[34]

Jeshion (2013a, 2013b) is one of the authors who resists the claim that stereotypes are semantically encoded in a given slur. She presents an argument where she pinpoints that one can utter a sentence like "Yao is Chinese." while expressing disgust and contempt. Her view is that the reason for the speaker expressing contempt is because they are committed to a stereotype about the Chinese. Clearly, the stereotype is not semantically encoded in the neutral word Chinese. The speaker evokes it via their attitude and expression of contempt; therefore, the stereotype does not need to be semantically encoded in a given slur. Another similar case where the neutral counterpart also has a negative connotation that comes to mind is from Croatia when referring to Serbs. There was a campaign in Croatia against hate speech and a commercial which listed certain slurs used in Croatia. While naming these slurs, among others such as *peder (Eng. faggot)*, *Cigan (Eng. Gypsy*[35]*)*, was also *Serbian*. The same goes for *Židov* in a sentence like "Ne budi takav Židov." (*Eng. "Don't be such a Jew."*) meaning "Don't be so stingy". This is not to say that there is a stereotype encoded in the neutral counterpart (a problem that Jeshion mentioned). The possible answer could be that for some groups, *Srbin* and *Židov* are used as slurs. This is not anything new; as Hughes (2006) writes, the word *whore*, for example, derives from Latin where its meaning is *dear*. The word picked up a negative meaning somewhere in the past and has been used negatively since. The same applied to the word *Negro* in the past when it was used neutrally. It was only later that it became a slur. The same approach can be applied to the cases where the neutral counterparts,

---

from our typical understanding of a (positive or negative) stereotype (for example, Chinese being technologically savvy, or them being bad drivers).

[34] However, it will be useful and needed at some point in the future to make this distinction. Therefore, I leave this for a future task.

[35] *Gypsy* here is a loosely translated term since *Cigan* in Croatian refers to the Roma people and always acts like a slur.

such as Srbin or Židov, are used in a negative way; they are being used as slurs. How is that possible? The explanation can be found, I argue, in the fact that slurs can be viewed as having *layers*. The notion of layers (or levels as Miščević interchangeably uses them) has been previously introduced by Miščević (2016), and I utilize this to introduce the negative identity prejudice layer, one that I find crucial for slurs. Although Miščević subscribes to semantic theories of pejoratives, as I announced previously, I will not be taking a stance on this. I feel that identity prejudice can be tied to slurs both in a semantic and a pragmatic sense. So, one of the said layers would be negative identity prejudice and, each time a neutral counterpart is used as a slur, it would mean that a new layer is attached to it, namely the negative identity prejudice we associate with being a Serbian or with being a member of the Roma population. This would also explain the reclaimed uses of slurs. When using the word such as *faggot* in a reclaimed sense, the layer that is the identity prejudice of being a homosexual is peeled off and replaced with a new, positive layer.

Another example presented in the literature as a case against stereotypes being semantically encoded in slurs was the case of *midget* which is inflammatory (Camp 2013). It seems to me that, even if there is no stereotype evoked by *midget*, there still is identity prejudice involved: identity prejudice of being a little person, which can entail, for example, being a freakshow or a spectacle. Furthermore, in a 2013 study by Jeremy D. Heider et al. it was found that adjectives associated with people with dwarfism were negative such as weird, childlike, incapable.

Thirdly, Jeshion (2013b) mentions examples such as the Yiddish *Goyim* used to derogatorily refer to non-Jews, or Japanese "'Gai-jin,' which literally means 'outside person'" (Jeshion 2013b, 322-323). She uses these examples to show how, since the words refer to all non-Jew or all non-Japanese people, there is no stereotype to draw from, but the words are still offensive. There is a similar term used in Croatia for all outsiders, *furešti*, which would refer to all non-natives or non-locals. But the thing is, even if we agree that there is no stereotype present in these cases, or, in fact, in the case of *midget*, there *is* a prejudice present, and a specific one at that: the negative identity prejudicial stereotype. It is a prejudice that refers to all out-group members who are less worthy, different, and therefore not as good as the in-group just because they are outsiders. This is part of our *conception* of them, i.e., we see this as part of their identity, and indeed, being an outsider *is* a part of their identity and the in-group members can hold identity prejudice over them.[36]

Another criticism about stereotypes being semantically encoded in a given slur comes from Mihaela Popa-Wyatt and Jeremy L. Wyatt. They are interested in the explanation of the

[36] I thank Miranda Fricker for discussing this with me and confirming that the broader conception of prejudice is a kind of resistance to counter-evidence caused by some affective investment.

offense slurs cause. They stress that "offence varies across different slur words, across different uses of the same slur word, and across the reactions of different audience members" (Popa-Wyatt and L. Wyatt 2018, 2880) and how "current theories struggle to parsimoniously explain the resulting patterns of offence" (ibid.). They explain that offense varies with the degree of oppression because many slurring utterances are oppressive speech (ibid.). To ensure a better explanation of their theory, they tackle how other theories of slurs view offense. Since they heavily rely on the explanation of offense in slurs and consider it to be an important matter in understanding how slurs work, I will take some time here to present their theory and provide some more insight before moving on. First, by using Hom's account of slurs as a prime example of semantic theories, they claim that semantic theories have some limitations in explaining offense. Hom is one of the authors who advocates that a stereotype is semantically encoded in a slur where derogatory force is tied to the content of the property it expresses and to the supporting racist institution (Hom 2008). Even though Hom and May (2018) have stated that the meaning of the slur cannot be offensive since the offense is a psychological property, Popa-Wyatt and L. Wyatt (2018) summarized the view in simplified terms, "offence can vary with word meaning because of semantic encoding of more or less negative stereotypes for different groups" (2884). They see this as problematic because, as they put it, "Hom's explanation cannot account so easily for variation across different slurs for the same group (intra-group variation). For example, 'nigger' is considered more offensive than 'spook'. But for Hom, in order for it to be so, it must be the case that the stereotypes encoded by the slur rest on two different racist ideologies" (Popa-Wyatt and L. Wyatt 2018, 2884), which cannot be the case because the institution for racism of the two terms should be the same. It seems to me that this issue could be solved by evoking the historical link the slur has. As presented by Miščević (2016), each slur carries a historical link in its semantics, which signifies the point in the past when somebody referred to the group by using a slur and the term stuck. We can claim that different historical settings of when the slur emerged and the significance it had in the past play a role in the perceived offensiveness. What determines the slur's force is not just the stereotype it carries, as it is also necessary to factor in the speaker's (non)culpability, the historical link, the setting in which the slur was used, the audience, and the speaker's intention.[37] Another, I suspect a more important reason I introduced earlier, is that we can view slurs as having *layers*. Layers in slurs can explain the appropriation process in the sense that when using an appropriated

[37] Perhaps a similar point is made by Davis and McCready (2020) where they say that the slur's expressive meaning component "imposes onto the context a complex of historical facts, stereotypes, and prejudices. This complex is not attitudinally linked to the speaker, and thus the conventional content of the slur does not entail anything about the speaker's attitudes to the invoked complex. However, the invocation of this content is unavoidably triggered by utterance of the slur; it is in this sense that we say the content is expressive" (Davis and McCready 2020, 9).

form, the "bad" layer of the slurs is peeled off[38] and replaced with a positive one. But, still, I would like to reiterate my point here that offensiveness is a subjective matter. As Hom and May also stated, offensiveness is not part of the meaning of a slur, and people may or may not be offended by stereotypes and prejudice directed at them. Luvell and Barnes (2022) made a point about this as well, noting how most literature about slurs only assumes offensiveness without providing any further elaboration of the term. There have been some accounts that provided more explanation of the term (Popa-Wyatt and L. Wyatt 2018; Bolinger 2017), but despite that, it seems that, perhaps, the notion of offense is too vague and broad to account for what happens when uttering a slur. A different notion may be better suited here, perhaps the one of the perlocutionary potential—the potential the slur has to degrade, offend, etc., its target. As Hom and May (2018) had already pointed out, offensiveness cannot be a part of the slur, but the slur's *potential* to do so perhaps can.[39]

In the above text, I have offered a brief overview of some charges that were made to encoding a stereotype semantically. The discussion above is just a preliminary discussion of the issues that is in no way conclusive. I think that the augmentation I am about to offer could work in both directions, i.e., whether stereotypes are encoded semantically or not. The key lies in the identity prejudice introduced by Fricker.

[38] I will say more about this later in the text.

[39] The notion of offensiveness and derogation is a tricky one and, as of recently, some authors have tried to challenge it (Liu 2021; Davis and McCready 2020). This issue is a one that merits a more thorough discussion (and a one worth having, I might add), however at this point, I am not prepared to delve into it.

3.5. Slurs and identity prejudice

In the previous Chapter, I have introduced the needed background on stereotypes and prejudice, as well as Fricker's notion of testimonial injustice. One of the central points in this thesis is that slurs evoke identity prejudice described by Fricker (2007). This merits a more thorough discussion. First, I will reiterate some points from the previous Chapter in order to say more about how Fricker understands identity prejudice.

Identity prejudice is a "label for prejudice against people *qua* social type" (Fricker 2007, 4). In other words, they are based on our social identities (what it is to be a man, woman, etc.). These are a kind of tracker prejudices, i.e., the ones that follow us through every social dimension, such as "economic, educational, professional, sexual, legal, political, religious, and so on" (Fricker 2007, 27). They can come in a positive or a negative form, but Fricker's interest, as well as mine, is primarily with the negative identity prejudice. Fricker explains that prejudice enters one's judgment "via stereotypes that we make use of as heuristic" (Fricker 2007, 30). Fricker is committed to Lippmann's (1922) description of stereotypes as images:

> If we think of a social stereotype as an *image* which expresses an association between a social group and one or more attributes, and which thereby embodies one or more generalizations about that social group, then it becomes clearer how its impact on judgment can be harder to detect than that of a belief with the same content. Images are capable of a visceral impact on judgment, which allows them to condition our judgments without our awareness, whereas it would take an unconscious belief to do so with comparable stealth. (Fricker 2007, 37)

She further states that stereotypes are "widely held associations between a given social group and one or more attributes" (Fricker 2007, 30).

Identity prejudices "typically enter into hearer's credibility judgment by way of the social imagination, in the form of a prejudicial stereotype – a distorted image of the social type in question" (Fricker 2007, 4). Moreover, identity prejudices enter one's judgment "often despite, rather than because of, their beliefs" (ibid.). She clarifies how the social atmosphere is one riddled with "stray residual prejudices" (Fricker 2007, 5) that may influence our judgment. These kinds of prejudices may be at work in the stereotype. Fricker defines prejudices as follows: "Prejudices are judgments, which may have a positive or a negative valence, and which display some (typically, epistemically culpable) resistance to counter-evidence owing to some affective investment on the part of the subject" (Fricker 2007, 35; removed italics). In fact, she elaborates how we should conceive prejudices as pre-

judgments, "where this is most naturally interpreted in an internalist vein as a judgement made or maintained without proper regard to the evidence, and for this reason we should conceive of prejudice generally as something epistemically culpable" (Fricker 2007, 33). By combining the two notions of prejudice and a stereotype, Fricker gets to the definition of a negative identity-prejudicial stereotype which she defines as follows: "A widely held disparaging association between a social group and one or more attributes, where this association embodies a generalization that displays some (typically, epistemically culpable) resistance to counter-evidence owing to an ethically bad affective investment" (Fricker 2007, 35; removed italics). Even though the discussion about judgment may, at first glance, seem to require consciousness, that is not necessarily so. Accordingly, Saul (2017) notes that Fricker's general commitment to a negative identity-prejudicial stereotype "does not seem to be one that requires consciousness" (Saul 2017, 236). In fact, Fricker herself notes how "prejudicial stereotypes can sometimes be especially hard to detect because they influence our credibility judgments directly, without doxastic mediation" (Fricker 2007, 36). What does that mean for our discussion on slurs? That means that the speaker may utilize negative identity prejudice even when they may not be completely consciously aware of it. A case in point would be the gendered slurs discussed in this Chapter. For example, when referring to a person as a *whore,* a speaker may believe they are only referring to an individual, as in a sentence like:

A: "Ana is a *whore.*"

The speaker may believe they have only said something about Ana, when, in fact, they are utilizing negative identity prejudice about women in general.[40] This is a crucial point. Even slurs that are typically not considered hate speech, as in the case of a slur *whore* or most gendered slurs,[41] produce various harms usually attributed to hate speech. But, as I will elaborate, these kinds of slurs, by evoking negative identity prejudice, produce various harms I will introduce in the next Chapter. In other words, the harm in the serious use of slurs comes from negative identity prejudice that is attached to a slur from the collective social imagination.

Having reiterated these notions about identity prejudice and how Fricker conceives this (and rightly so), I will claim that these negative identity prejudices "float" around in the social imagination; they follow us through every social dimension and thus also stay with us in a discourse setting. When the slur is uttered, they *stick* and *attach* themselves to a slur and are evoked every time a slur, in its literal sense, is uttered, i.e., when the speaker uses a slur

[40] The case for why that is so is explained earlier in the text and will be further elaborated on in the later text, as well.

[41] I am indebted to Enes Kulenović for pointing this out to me.

in its literal sense to degrade, as a weapon, i.e., where the use is intended, slurs attach negative identity prejudice and cause harm. Usually, this is done by competent speakers who know what they are doing. The most harm is done, of course, if the use of a slur is systematic. To paint a very vivid picture, these prejudices would function as viruses: each time a slur is uttered,[42] identity prejudice attaches itself to it. In that sense, we can say that slurs become a sort of an embodiment of identity prejudice. Also, a lot of the cases of such uses of slurs will be directed at historically marginalized groups since identity prejudice usually (but not always and in no way exclusively) targets historically marginalized groups.

The sticking and attaching of negative identity prejudice can be understood in a semantical or a pragmatical sense. As stated numerous times, I will not discard either possibility here and this decision warrants more exhaustive research. The central point here is that when uttering a slur, we give a negative normative evaluative judgment about the target. This normative evaluative judgment is fueled by negative identity prejudice. Fricker (2007) sees prejudices as judgments (or, more precisely pre-judgments) and by uttering a slur, the speaker normatively judges the target in a negative way. Even though when Fricker (2007) talks about judgments, she focuses on credibility judgments, we can extrapolate this view to encompass a wide array of normative judgments. To turn to slurs, in an already mentioned example sentence:

A: "Ana is a *whore*.",

the negative identity prejudice that women should be ashamed of their sexuality and should aspire to appear lady-like, grounds the evaluative judgment of the target (in our case Ana), i.e., that the target is bad. As stated, negative identity prejudice (which Fricker views as pre-judgments) can influence our judgment consciously or unconsciously. As noted before, by Saul (2017) and by Fricker (2007), the identity-prejudicial stereotype does not require consciousness. By using slurs, we directly make use of and activate the identity prejudice. The hearers do not need to endorse what the speaker is saying—in fact, they may even challenge the speaker. However, when the speaker employs negative identity prejudice, this may trigger some unconscious processes in the hearer even when the hearer is not aware of it due to the stealth mode identity prejudice can take. So, even though the hearer may not agree with the speaker, the utterance may still unconsciously influence them.

Furthermore, identity prejudice gives us a plethora of prejudices to take as a central feature depending on the context. In a sentence like:

A: "Of course he failed the course, he is a *Ni**er* after all!"

[42] By uttered I mean used, mentioned, written, etc.

what we take as central is one's intellectual ability, i.e., the identity prejudice we draw from is that African Americans are intellectually inferior. Which identity we choose to focus on would depend on the context.

Let me now turn to a few possible problems. Some of them have already been mentioned.

I have already mentioned the example of two slurs that supposedly evoke the same racist stereotypes, but their offensive potential is starkly different. Namely, *spook* and *ni**er* have different offensive potentials where the former is not as offensive as the latter. As already stated, the force of a slur is not determined *just* by attaching identity prejudice to it. A slur's force can be weakened or strengthened by other features, such as the historical layer, i.e., who said or coined the term and the context in which it was used in the past, etc. Slurs may share the same, or nearly the same, negative identity prejudice and still differ in the force they exhibit. Secondly, Jeshion (2013a) gives an example where someone may utter the slur *Chink* and do so while only feeling contempt for the target based on their ethnicity, not knowing anything about stereotypes usually associated with the Chinese. As I understand this example, the contempt would come from nothing more than *otherness*, the fact that they are different than the speaker. I understand identity prejudice to be more along the lines of what Jeshion (2013a) described to be an identifying component of the target. In the case of *faggot*, she explains: "That is, it follows from what it is to find someone contemptible on the basis of being gay that one takes that person's sexual orientation as the most or among the most central aspects of that person's identity" (Jeshion 2013a, 318). So, the identifying component of being gay is tied to one's sexual orientation. A crucial point to remember is that negative identity prejudice is tied to our social identity (what it is to be gay, to be a woman, to be black, etc.), whatever that entails in a given context. Therefore, in the case of *faggot*, the negative identity prejudice would be the identifying component of being gay, which is that gays enter into same-sex relationships which is bad because it is unnatural and only different sexes should get romantically involved.[43] In the case of *Chink* where the speaker expresses contempt based solely on the ethnicity of the target, the identity prejudice would again be tied to whatever social identity we take as central in a given context. In this case that would be being of Chinese origin which would mean being culturally different from the speaker. Jeshion (2013b) correctly notices that some of the cases of slurring can be based solely on perceiving others as different. Indeed, in some cases prejudice can only be based on being different than the in-group. That is the case with the Croatian slur *furešti* used to refer to all non-natives or non-locals. Sometimes we do hold prejudices against foreigners solely based on them being in the out-group, meaning not being as good as the in-group,

[43] Cue the similarity between the case of *faggot* and gendered slurs where the identity prejudice that women should not exhibit P-behavior fuels the slur.

being different in some bad sense. Perceiving someone as being bad because they are in some way different than us, even when we sometimes can't pinpoint what that difference amounts to, means having a prejudice against them. In that case, we take their being part of the out-group as a central feature of their social identity. In other words, we hold identity prejudice over them. It is also crucial to remember that, in Fricker's view, a negative identity prejudicial stereotype means being resistant to counterevidence while having some affective investment.

Another possible concern are cases of non-intended uses of slurs, i.e., where the speaker makes a non-culpable mistake, such as in the case of a non-native language speaker. In that case, the speaker would not hold a prejudicial stereotype although it could be the case that the audience may attach the identity prejudice to what was said. That's why some may even take offense. However, the speaker could easily backtrack, and this case would amount to an honest mistake. There are some slurs that are especially tied to negative identity prejudice almost in all cases, such as the N-word, so much so that even mentioning such a word may cause offense. But, even in those cases, the speaker can make an honest mistake and backtrack after being presented with counterevidence so we cannot say they hold a negative prejudice against black people. Nonetheless, since the connection between these kinds of slurs and negative identity prejudice is a strong one, the audience may *presuppose* that the speaker holds such prejudices and might take offense, at least until the issue is cleared and it becomes evident the speaker made an honest mistake.

The last thing to address would be the question related to the reclamation of slurs, namely, why can only in-group members use a slur among themselves while the same is not the case for out-group members even when the out-group member intentionally doesn't attach the identity prejudice to a slur? The potential answer to this question could be found in Fricker's relation of power in identity prejudice. Fricker (2007) introduced identity power as a subset of a broader concept of social power where identity power "is a form of social power which is directly dependent upon shared social-imaginative conceptions of the social identities of those implicated in the particular operation of power" (Fricker 2007, 4). In that sense identity power (or any other kind of power) the out-group member holds would infringe on their use of a slur even in cases where the out-group speaker would purposefully intend to detach identity prejudice from the slur and use it in its reclaimed form. These power relations in a discourse would then dictate who can use slurs and in which context they can be used.

Now, in order to construe how identity prejudice could possibly fit into a semantic account of slurs, I will utilize and adapt Miščević's (2016) idea of slurs as having levels/layers. I reiterate here once again that I will not take a stance on whether the identity prejudice works as a semantic or a pragmatic device, or whether any of the layers could also

be part of the pragmatics of slurs. In fact, I am open to an interpretation where any of the layers could fit into a pragmatic perspective.

The notion of layers (or levels, as Miščević interchangeably uses them) has been previously introduced by Miščević (2016). I utilize this to introduce the negative identity prejudice layer, one that I find crucial for slurs. Miščević subscribes to semanticist theories of pejoratives, and I will adapt his layered approach to slurs in order to account for the identity prejudice layer. However, as stated before, I think that identity prejudice can be tied to slurs both in a semantic and a pragmatic sense and I am open to the idea that some of the layers that Miščević holds to be semantically encoded, could be pragmatically conveyed. Furthermore, for what I am about to claim later on, the issue of whether the identity prejudice is semantically encoded or pragmatically processed is not essential at this point, although it will warrant future disentanglement.

We have seen from the example of the slur *bitch* presented in this Chapter that there are several different meanings for the same word because the word has changed and evolved. This feature of how the same slur can have different meanings is perhaps best explained by understanding slurs to have layers, and I portray this with the table below (the table is an adjusted form of Miščević's example from 2016). Thus, taking into account the connection drawn between slurs and Fricker's account of identity prejudice in testimonial injustice, we the issue with slurs can be portrayed as follows:

Table 2

A: Ana is a *whore*.	**Layer**
	1) Neutral counterpart: Ana is a (promiscuous[44]) woman 2) Negative normative evaluative judgment: Ana is a bad person and should be avoided because she exhibits (some) negative characteristics and that is bad because she is a woman 3) Negative identity prejudice: women should be ashamed of their sexuality and exhibiting promiscuous behavior is not lady-like—which is something a woman should aspire to be, exhibiting some kind of sexual behavior devalues a woman's worth in society 4) Historical link: at some point in the past the target group was labeled with the slur 5) A feeling of contempt 6) Epistemic (non)culpability

The table above[45] is just a very broad conception of how slurs would work. I have left the left side of the table empty, i.e., there is no semantic/pragmatic distinction as in Miščević's example. The reason for this was previously mentioned and elaborated on. Namely, for the purpose of introducing derogatory-labeling injustice, this distinction, even though rightfully important for any theory of slurs, is not crucial at this point. I am aware that, for a theory of slurs to be complete, it needs to provide an explanation of a slur's pragmatic and semantic elements. However, for the purpose of this thesis, this is not a crucial point since derogatory-labeling injustice could happen in both cases: whether negative identity prejudice is semantically encoded or pragmatically conveyed. What is a central point is that the negative evaluative judgment is grounded in negative identity prejudice. Introducing negative identity prejudice provides us with an explanatory advantage of a slur's content: the negative evaluative judgment is grounded in negative identity prejudice, i.e., the

[44] See the explanation in this Chapter of how I understand the term promiscous and that it actually refers to P-behavior.

[45] The table obviously borrows and incorporates notions and explanations provided by authors who support semantic theories (such as Miščević (2016) and Hom (2010)). As said before, I will not argue for semantic theories here, although I am sympathetic to their cause. However, I think the identity prejudice view could work even if the identity prejudice is placed outside the literal meaning of a slur.

negative identity prejudice layer explains why slurs have a negative evaluative judgment layer. Or, to be even more bold, the negative evaluative judgment exists due to negative identity prejudice.

So, let us portray the aspect of layers with another example:

a) X is a *ni**er*[1].

In a) the *ni**er*[1] would signify a literal meaning where the speaker conveyed that X is:

layer 1) a black person,

layer 2) X is a bad person and should be avoided because he exhibits (some) negative characteristics due to being black,

layer 3) negative identity prejudice that black people are lazy, unintelligent, violent, prone to anger, and so on, because of being black,

layer 4) historical link to the time members of the target group were labeled with the slur,

layer 5) a feeling of contempt and disgust towards members of the target group,

layer 6) epistemic (non)culpability.

Layer 1) serves as the neutral counterpart of the slur.[46] The second layer provides a negative evaluative judgment about the target, and it ascribes negative characteristics to the target. Notice that in Table 2 and the above example, layer two is described differently: in the case of *whore*, layer two states that the target exhibits (some) negative characteristics and that is bad *because* she is a woman, whereas in the example of *ni**er*, layer two states that the target exhibits (some) negative characteristics *due* to being black. But, this shouldn't be problematic. As explained earlier in this Chapter, in the case of gendered slurs, the issue is with exhibiting P-behavior which is viewed negatively because one is a woman, and women shouldn't exhibit such behavior. In the case of ethnic slurs or racial slurs, it is usually the case that the speaker attributes certain behavior to the target's race or ethnicity, for example, he is violent because he is black. In each case, however we construe the evaluative judgment of the target, a certain feature of negative identity prejudice was taken as central. The third layer, or the negative identity prejudice layer, explains the second layer: the negative evaluative judgment is grounded in the negative identity prejudice. The speaker makes an evaluative judgment of the target based on an identity prejudice they take to be a central

[46] There are accounts that challenge the assumption of a neutral counterpart (see Ashwell 2016), however, I am inclined to keep this distinction even for more dubious slurs, such as gendered slurs, and I elaborated on the reason for this earlier in this Chapter.

feature of a target's identity depending on the context. The historical link layer, the fourth one, can help us understand why some slurs are perceived to be more offensive than others, as was described in the case of *spook* and *ni**er*. In the fifth layer, the speaker expresses their emotional attitude towards the target. The sixth layer comprises the speaker's epistemic (non)culpability, a term borrowed from Fricker (2007), in which we can consider whether the speaker is culpable for uttering a slur. This layered account could help us explain some non-intentional uses of slurs, or to put it in Fricker's terms, non-culpable uses. The uses I have in mind here are uses such as when the speaker is a non-native speaker and is unfamiliar with the fact that a certain word they uttered is a slur. In that case, layer 5) would be erased: the speaker doesn't feel contempt towards the target since they are completely unaware that they are uttering a slur. I will elaborate more on layer 6) in the fourth Chapter, where I will make a connection between slurs and Fricker's epistemic injustice.

The above proposal is, of course, in need of much more in-depth analysis and research but it is perhaps a path forward to a new way of thinking about slurs.

In this Chapter, I have introduced the theories of slurs in order to understand how they work from a perspective of the philosophy of language. I have mostly focused on stereotypes that are evoked when uttering a slur, and I have introduced a novelty in literature about slurs and stereotypes—namely, that the stereotype evoked by slurs is a negative identity prejudice first described by Fricker (2007). However, in order to fully understand how slurs work, one must take into account their potential effects on society. This is a task I will be focusing on in the next Chapter. Utilizing the research from the previous Chapters, I will turn to the central case of this thesis—the introduction of a novel notion of derogatory-labeling injustice.

CHAPTER IV: THE EFFECT OF SLURS AND DEROGATORY-LABELING INJUSTICE

4.1. Introduction

After setting up a needed background, I turn to the pivotal aspect of this thesis—the introduction of a novel concept of *derogatory-labeling injustice* (inspired by Fricker's and Kukla's notions) which emanates from the systematic uses of slurs. Derogatory-labeling injustice happens when the speaker, who is in a position of power, labels the target with negative identity prejudice by using a derogatory word, i.e. slur, and thus produces one or more harms to the target's interests. This kind of injustice hasn't been described in the literature so far, but it was foreshadowed by Fricker (2007) when she claimed that one is susceptible to "a gamut of different injustices" (Fricker 2007, 27) due to tracker prejudice she identified. This tracker prejudice, I claim, is evoked by slurs and is responsible for various harms inflicted on the target. Even though the claim that slurs harbor prejudice and stereotypes isn't a novelty in the literature, the claim that these stereotypes and prejudice are of a specific kind—namely that the stereotype in question is a negative identity prejudice identified by Fricker—is. I will claim that due to the negative identity prejudice they harbor, slurs enact social harms on their targets. Slurs producing harm has mainly been only presupposed in literature and not much systematic work has been done. In other words, focusing on and tracing the harm back to stereotypes and prejudice has, to my knowledge, been scarce, especially when focusing on slurs. Thus, my contribution to this is that I use empirical evidence on stereotypes and prejudice in order to pinpoint exactly what kind of harm slurs do. I examine the effect of slurs from three perspectives: *the speaker*, *the target*, and *the listener* since the debate about the harm slurs do has to consider all of the said perspectives in a discourse to grant a full understanding of what happens when a slur is uttered. Additionally, when discussing harm that has a long-lasting effect, I examined some harms that haven't been discussed in literature before, namely the ability to impede opportunities to acquire primary goods and hinder thinkers' interests (as an answer to Seana Shiffrin's thinker-based account).

Furthermore, it seems that the concepts of hate speech and slurs have so far not been able to fully explain the harm slurs might do. Namely, I've previously explained that slurs are a vehicle of hate speech, however, it seems that slurs can do harm even when they are not considered hate speech. This is where the novel concept of derogatory-labeling injustice is introduced to grasp the full possible extent to which slurs might harm, even in cases that are not considered hate speech.

So, my approach in this Chapter will be as follows: first, I will aim to make a connection between Fricker's testimonial injustice and slurs and then review what slurs do when uttered. I will review this from the perspective of the speaker, the listener, and the target. I will claim that slurs enact social harms on the target and that these harms can be manifested in two effects: the primary and the secondary effect, where the primary effect is more immediate, and the secondary effect is more consequentialist in the sense that it may have a long-lasting effect. As Bonotti and Seglow (2021) mention, there have been some recent worries put forth in the literature that arguments that claim hate speech has a long-lasting effect have a hard time tracing the source of these particular effects back to hate speech since the said effects can arise from other injustices. Therefore, some scholars (Simpson 2019; Heinze 2016 as cited in Bonotti and Seglow 2021) maintain that authors who argue that hate speech produces a long-lasting effect bear the burden of proof and should back up their claims with empirical research. I partly agree with this claim. My claim is that there is a specific prejudice evoked when a slur is used in its literal sense to degrade—it is the identity prejudice described by Fricker. Thus, in Chapter II, I have provided empirical research on stereotypes and prejudice, and even on effects of derogatory language, which provides a needed background to claim that the harm done by slurs stems from stereotypes and prejudice. Slurs make up a large portion of hate speech, but, as explained in Chapter I, hate speech can take many forms. I think that the view that stereotypes and prejudice fuel slurs can be extrapolated to understand other aspects of hate speech as well. Therefore, the background provided in Chapter II can be of use to understand how hate speech manages to enact such social harms. After presenting the effect of harm caused by slurs, I turn to the pivotal aspect of this thesis—the introduction of *derogatory-labeling injustice* (inspired by Fricker's and Kukla's notions) which is produced by systematic uses of slurs. I leave the introduction of derogatory-labeling injustice for the very end since, in order to grasp the concept of it, one must first understand the scope and the mechanism of harm being produced.[47]

[47] Similar method was argued by Kulenović (2023): "the assumption is that within political theory the content, scope and true character of hate speech itself is often determined by our understanding of why this type of speech is dangerous and why it should be curtailed" (Kulenović 2023, 512).

4.2. Slurs and testimonial injustice: the connection

I believe that Fricker has correctly pinpointed the stereotypes and prejudice that are at work in society, or as she puts it, that preside in our collective social imagination. To reiterate, stereotypes are something we make use of from our earliest days as a kind of shortcut to make sense of the world around us and to make our lives easier by categorizing. We use them on a daily basis in our communication whether we are aware of it or not. The problem is, as Fricker correctly notes, when prejudice enters the picture. The prejudice Fricker is concerned with is identity prejudice: the kind of prejudice that is based on our social identities (what it is to be a woman, gay, and so on) and that follows us through every social aspect of our lives. As she elaborates, a stereotype can harbor identity prejudice, and identity prejudice is in many cases connected to historically marginalized groups, for example, women, people of color, gays, etc. Furthermore, when we make a judgment about people, we can make a non-culpable mistake if we correct our belief system when we encounter counterevidence. But, if our belief system is not adjusted when encountered with counterevidence, then we can say that we harbor prejudice, and not just any kind of prejudice, but the one Fricker calls a negative identity-prejudicial stereotype which she defined as follows: "A widely held disparaging association between a social group and one or more attributes, where this association embodies a generalization that displays some (typically, epistemically culpable) resistance to counter-evidence owing to an ethically bad affective investment" (Fricker 2007, 35).

I believe this is the exact kind of stereotypes and prejudice that are at work in slurs. Furthermore, I think advocating for such specific prejudices and stereotypes evoked by slurs gives us an explanatory advantage when trying to grasp what happens when slurs are uttered. Let me elaborate.

First, in order to make the connection between slurs and testimonial injustice, I will utilize Fricker's idea of social power where social power is "a practically socially situated capacity to control others' actions, where this capacity may be exercised (actively or passively) by particular social agents, or alternatively, it may operate purely structurally" (Fricker 2007, 13). From this conception of social power, Fricker then postulates an idea of identity power where agents have shared conceptions of social identity, namely "conceptions alive in the collective social imagination that govern, for instance, what it is or means to be a woman or a man, or what it is or means to be gay or straight, young or old, and so on" (Fricker 2007, 13). Fricker provides an example of gender functioning as one domain of identity power where the active use of identity power would be a man using "his identity as a man to influence a woman's actions—for example, to make her defer to his word" (Fricker 2007, 14), and the passive use of identity power would be when a woman is silenced "by the mere fact that he is a man and she a woman" (Fricker 2007, 15) in which case the man doesn't

have to actively do anything. Identity power depends on imaginative social co-ordination: "both parties must share in the relevant collective conceptions of what it is to be a man and what it is to be a woman, where such conceptions amount to stereotypes (which may or may not be distorting ones) about men's and women's respective authority on this or that sort of subject matter" (Fricker 2007, 15). We can postulate that every discourse has a certain power relation between the speaker, the listener, and the target, bearing in mind that sometimes the listener and the target are the same person and sometimes they are not. Each person joins the discourse bringing with them a certain amount of power. For example, in the discourse setting between a man and a woman, a man has identity power over a woman; or, one can be in a position of a different kind of power, such as hierarchical, political, and so on. When the speaker is in a position of power, the perlocutionary potential the slur has could be even greater. For example, due to their social role, politicians have political power which allows them to reach and potentially influence a greater number of people. We have witnessed an example of the perlocutionary effect a politician's word may have during the January 6 United States Capitol attack in 2021 when a mob of Donald Trump supporters charged the Capitol following his defeat in the elections. Trump was the one who, by giving a speech in which he repeated some false claims about the election, (willingly or not) encouraged his supporters to act. Later on, he further encouraged them by writing on social networks. This all led to rioters entering the Capitol building.

Second, slurs target groups, and when we utter a slur, we give a normative evaluative judgment about the target. However, it seems that there is some confusion in the literature about what words would fall under the category of a slur. For example, Nunberg (2018) has claimed that there are no slurs for women since gendered slurs target only individuals.[48] Bach (2018), for instance, distinguishes between group slurs (such as *kike*) and personal slurs (which Hom (2010) would consider insults). Viewing slurs as an embodiment of identity prejudice can solve some of these issues. Identity prejudice, as per Fricker, usually (but not always) targets historically marginalized groups. By tying identity prejudice and slurs, we can extrapolate that slurs also mostly target historically marginalized groups. Slurs that produce serious harms I will mention below, are slurs that refer to historically marginalized groups, and only in that case could a slur produce derogatory-labeling injustice. Furthermore, the normative evaluative judgment of the target is actually fueled by the identity prejudice Fricker described. Of course, identity prejudice is not in any way exclusively tied to historically marginalized groups, they can and do refer to dominant groups as well. However, my focus here is specifically on historically marginalized groups since, due to their fragile position in society and due to the already existent prejudice, they are more susceptible to various harms slurs may inflict on them. Because of this, I feel that derogatory language

[48] I have already provided a possible explanation of this view in Chapter III.

directed at these groups can be especially harmful and thus needs to be addressed accordingly to ameliorate the potential harms.

Let me now portray how negative identity prejudice is evoked by considering some examples.

In the dictionary the word *whore* is described as having two meanings, one is a female prostitute, and the other is a woman who sleeps with a lot of men. The word has a negative valence which is rooted in the negative identity prejudice of women who should be ashamed of their sexuality and not be open about it because that is not lady-like (something all women should aspire to be). Examples from real-life situations that depict this are in abundance. It's common knowledge that men who have a lot of female sexual partners are viewed as experienced and it is considered to be desirable; in fact, men who don't have sexual experience are viewed negatively. The opposite is true for women. In fact, women don't even have to have a lot of partners, they just need to express their sexuality more openly (for example by what they wear) to be branded a *whore* or a *slut*. Let us remind ourselves of a case of a Canadian police officer who warned women that they might be to blame for sexual assaults if they are not careful with how they dress.[49] This view that slurs harbor identity prejudice applies to other slurs as well. For example, a slur *ni**er* harbors negative identity prejudice of being black (prone to violence, less intelligent, and so on).

So, it would seem that identity prejudices that linger in the collective social imagination, and as Fricker notes, that follow us through every social dimension, have found their embodiment—in slurs. Since identity prejudices follow us through every social dimension, they also stay with us in a discourse setting. By using slurs, we directly make use of and activate the identity prejudice. Tying identity prejudice to slurs gives us an advantageous position: we are now better able to understand why slurs are perceived to be so powerful with such a potential to wound. It is because they evoke identity prejudices that linger in the collective social imagination and poison the well of a delicate ecosystem that is a community of heterogeneous people.

Moreover, slurs create an atmosphere that is fertile ground for testimonial injustice to take place. Let me portray this with an example of a person who holds a position of power[50] (suppose he is the CEO of a company X) and who uses a slur to refer to his female colleagues (suppose he refers to them as *whores*) in absentia. Since, let's imagine, he is addressing his male subordinates while calling his female colleagues *whores*, we can suppose that he

[49] For further reference read about SlutWalk, a movement that was inspired by this event.

[50] By the term "position of power" I understand "power" in Fricker's sense, i.e., identity power, but I also mean hierarchical power exercised through a person occupying, for example, an executive position in a company.

receives credibility excess[51] from his audience. In a society where stereotypes and prejudice make up a normal part of our daily lives, and where they can linger on in our minds without us being aware of them, the described exchange between an executive and the subordinates can increase the chances of testimonial injustice happening. The audience, being in a subordinate position, might feel pressure to agree with the speaker and to trust him in his characterization of their female colleagues. Perhaps in the next exchange with their colleagues, they will take the colleagues less seriously. In other words, it could be the case that the target's knowledge-status will be affected in the workplace since others may start doubting her expertise and start believing that she was hired, not because of her knowledge, but because of other, less respectable means (Perhat 2016). In that sense, slurs contribute to testimonial injustice and also, since evoking identity prejudice, they spread and sustain prejudice about targets.

But, by using slurs the speaker not only evokes stereotypes and prejudice, they also do something else. To borrow from the terminology proposed by Fricker, we can suppose that the speaker using slurs does something that is considered to be epistemically culpable. It is theoretically possible that the speaker using a slur is unfamiliar with the slur (maybe they are using a slur that is not in their first language or it's a new word for them) and they are unfamiliar with the slur's meaning. In that case, if the speaker is presented with the real meaning of the slur, we would expect them to stop using it in which case what they did when using a slur was an honest mistake, a non-culpable mistake. Or, the speaker might be using a slur for a purpose different than degrading the target, for example jokingly or among friends.[52] But, in cases of the literal use of slurs, the use where the goal is to degrade the target, where the speaker is very aware of the meaning of the slur, and where attempts at providing counterevidence have failed and the speaker didn't adjust their belief system accordingly, we can say the speaker, as Fricker puts it, does something epistemically culpable. Moreover, they do something ethically and epistemically bad, to borrow from Fricker again. Considering all of the above, when the speaker uses a slur and thereby evokes a negative identity prejudice, and when the speaker is doing that with malicious intent, i.e., they are epistemically culpable, we can claim they are doing something ethically bad. The speaker, by using slurs, is causing *harm*. The harm done can be immediate, or it can echo into a long-lasting effect. Actually, by taking all of this into account we can say that the speaker is engaged in something I will refer to as *derogatory-labeling injustice*. To understand what exactly this kind of injustice is, it is best to first shed more light on how

[51] Credibility excess occurs when the speaker receives "more credibility than she otherwise would have" (Fricker 2007, 17). In this case the executive of a company would possibly receive credibility excess from his male subordinates due to holding a position of hierarchical power.

[52] The appropriated uses of slurs are a separate topic and will be discussed in later chapters.

slurs work, how exactly they cause harm, and the role the speaker, target, listener, and society play in all of this.[53]

4.2.1. The harm done via slurs

Now that we have finally established a needed background in terms of presenting empirical evidence about how stereotypes and prejudice work and presented Fricker's case for testimonial injustice, we can move on to the central case of this thesis: explaining what exactly is "the bad" in slurs, and, consequently, hate speech, or, in other words, what is the harm the speaker is doing when engaging in hate speech via slurs.

However, before moving on, there is an important issue of demarcation that needs to be resolved. Namely, in the following text, I will mention hate speech and slurs and it will sometimes seem I use them interchangeably. That is because I will extrapolate the arguments given by authors for the harm done by hate speech—to slurs. In other words, my claim is that the arguments I discuss in this thesis that consider the harms done by hate speech could be applied to slurs. So, when I discuss harm produced by hate speech, slurs will necessarily be included in this equation. Let me elaborate on this a bit further. In some cases, there will be an overlap between hate speech and slurs, meaning that some slurs are considered hate speech. The obvious examples would be referring to Jews as *kikes* or to black people by using the N-word. Furthermore, some of these examples could also fall into the category of *fighting words*, the doctrine mentioned in the first Chapter. Of course, *fighting words* are always context dependent.[54] Slurs I am concerned with here, and that can produce harm, are slurs used for degrading the target (so, jokes, sarcasm, and similar are left out) and, in most cases, they can be considered hate speech. However, there are cases where slurs could produce some

[53] It is probably worth mentioning at this point that Popa-Wyatt and L. Wyatt (2018) could have an account that could be viewed as a complimentary one that could fit in, at least partly, with Fricker's notion of social and identity power. Namely, their account holds that "slurring utterances seek to create (or maintain) an unjust power imbalance via role assignment. Our second contention is that the degree of offence caused is correlated with the magnitude of the perceived unjustness of the power imbalance associated with this role" (Popa-Wyatt and L. Wyatt 2018, 2888) where "roles are social constructs that carry information about permissible and expected behaviors, social status (i.e. rank relative to other roles), rights, and responsibilities" (Popa-Wyatt and L. Wyatt 2018, 2888). To this, they introduce the notion of discourse roles where discourse roles adopt different social roles depending on the context. When using a slur, the speaker can put himself in the position of power and assign the target the subordinate role. In that sense, slurs contribute to oppression (Popa-Wyatt and L. Wyatt 2018). But, the assigned discourse roles also influence social roles in the real world beyond the discourse and this is achieved by perlocutionary effects of slurs. Summarily, "perlocutionary effects causing oppression beyond the discourse include: emotional injury, implicit threat of physical injury, silencing, increased permissibility and/or pressure for other oppressive acts, and increased desire to act oppressively so as to gain power" (Popa-Wyatt and L. Wyatt 2018, 2898).

[54] See Chapter I in this thesis.

of these harms, but would not fall into the category of hate speech. This issue can be portrayed in the following manner:

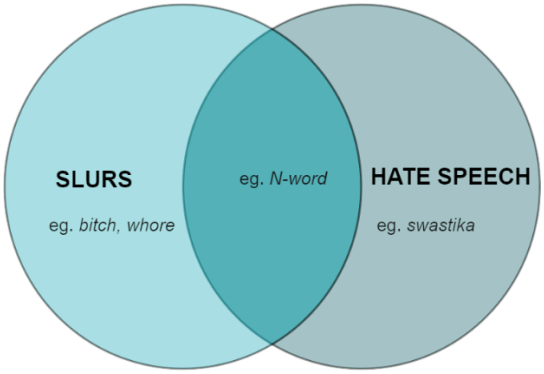

Figure 1

I view hate speech as a broad concept incorporating acts (such as cross burning), or symbols (such as the swastika), as well as words. The most used (but not exhaustive) vehicle of hate speech is slurs. Thus, there are some slurs that would be considered hate speech by most courts and laws. However, there are slurs that wouldn't be considered hate speech by most courts and laws, and the prime examples of these slurs are gendered slurs for women such as *whore* or *slut*. It seems that these slurs, when used in the literal sense to degrade, also cause harm to its targets. Thus, I will aim to show how these slurs that would not be considered hate speech should also be addressed because they too produce harm that is usually accredited to hate speech. More specifically, slurs produce what I refer to as derogatory-labeling injustice, a novel form of injustice that needs to be considered alongside hate speech. What is important is the fact that the most serious harm that would warrant possible legal action would be the use of slurs as hate speech. However, the harms I will discuss below can also be produced by slurs that would not be considered hate speech. I will say more about this once I introduce the notion of derogatory-labeling injustice.

We can, at this point, review what was said about slurs and how they work. It seems slurs function as having *layers*. These layers also help us understand the multiple meanings of slurs listed in dictionaries, as well as the appropriated uses of slurs. In the third Chapter, it was said that slurs are a vehicle of hate speech and that they express derogatory attitudes (Hom 2010). To reiterate, semanticist theories of pejoratives claim that slurs hold most of their content in their semantics. For example, every slur has a neutral counterpart, such as *gay/homosexual* for *faggot,* and non-semanticist theories of pejoratives tend to claim that (only) the neutral counterpart of the slur (in this case *gay*) is part of the meaning, and the negative part we are conveying when uttering a slur is all part of pragmatics. But, semanticist

theories don't agree. According to the semanticist theories, in addition to the neutral counterpart of a given slur, the meaning of slurs is loaded with the negative content slurs carry. So, the negative evaluative content of a slur would be a part of semantics, a part of the very meaning of the word. In the case of uttering a sentence like "Ana is a *whore*" what the speaker is (roughly) conveying by the meaning of the slur *whore* is that Ana is a bad person with a questionable moral compass who should be avoided all because of being a promiscuous woman. All of this is encoded in the meaning of the slur. The feeling of disgust the speaker is (probably) also conveying would be a matter of pragmatics. Non-semanticist theories claim that slurs are best explained by pragmatics and they place only the minimal content in the semantics of a slur. However, a number of both semanticist and expressivist authors claim that there is a stereotype evoked by a slur; however, they disagree on the placement of it—whether it resides in the semantics or the pragmatics. As already explained, I will not be taking a stance on this. My aim is to augment existing theories by proposing that the stereotype in question is actually the negative identity-prejudicial stereotype proposed by Fricker. Identity prejudice may work for both semanticist and expressivist theories. So, what happens when a slur is uttered, and how does negative identity prejudice affect us? The answer to this is complex, and even though much was said in the literature about slurs and how they work, as well as hate speech in general, not much work has been done (at least to my knowledge) to systematize the empirical evidence found on the effects of prejudice and stereotypes and how they apply to slurs. Slurs that evoke negative identity prejudice produce various harms that have an effect on our social lives. In order to fully grasp the scope of this, and to better understand what slurs do once uttered, it is best to review these effects from various perspectives: the speaker's, the listener's, and the target's. Each perspective has its role in the discourse setting where derogatory-labeling injustice takes place.

4.2.2. The speaker

First, let's consider the speaker, and let's start by reviewing the issue of culpability, inspired by Fricker. We can consider the speaker to be either epistemically non-culpable or epistemically culpable. The non-culpability is identical to an honest mistake due to the lack of knowledge on the part of the speaker. As previously explained, this can happen when, for example, the speaker is unfamiliar with the slur they use and therefore with the prejudice the slur carries. Although we can claim here that it is desirable to be epistemically responsible agents and to do our homework regarding the social community we are a part of, honest mistakes can still happen. What is important, and what would render something to be an honest mistake, is the response of the speaker once they are introduced to the meaning of the word. If the response is that they apologize and accept the explanation, then we may consider

them to be non-culpable. There are cases, of course, where we would not need to be so forgiving. That would be the case of, for example, politicians who claim they did not know the meaning of a slur they have used. Considering their social and political role, it is almost a requirement for them to be familiar with such things. We can consider the speaker to be epistemically culpable in cases where they refuse to change their belief system[55] once they are informed of the real meaning of the slur. In addition to that, the perlocutionary effect of an utterance will be greater if the speaker occupies a position of power, be it identity power Fricker described (Fricker gives an example of gender acting as one arena of identity power where the active use of identity power would be a man using "his identity as a man to influence a woman's actions—for example, to make her defer to his word" (Fricker 2007, 14), and the passive use of identity power would be when a woman is silenced "by the mere fact that he is a man and she a woman" (Fricker 2007, 15) in which case the man doesn't have to actively do anything), or other powers, such as economic or political power. As we have seen, hierarchical power in the example I have provided earlier in this Chapter brings with it a certain amount of pressure on the subordinates to agree with their superior, somebody they might even give credibility excess to. Power such as political one provides the speaker with a greater audience and the ability to reach more people through media and the like. The same would actually apply to anyone who has the means to reach out to a greater number of people, but politicians, in particular, are at even greater scrutiny since they are elected representatives and, through their role as politicians, they represent not only the voices of the ones that elected them but also the voices of an entire electorate, i.e., they represent their country. In that sense, there can be degrees of culpability where we can hold some people more accountable than others due to the social power they hold. These various powers can determine how far-reaching the effect of the slur may be.

So, when uttering a slur (in its literal sense) three things happen:

1) First, the speaker is attempting to degrade the target. By degrading I mean to view the target as "less than" in a sense the target is portrayed as not an equal member of the community, i.e., the target's social status is eroded. As I will understand it here, degrading will signify a type of action that is on the one hand psychological in nature, and on the other practical in nature. The psychological part is to conceive, to think about other people as less worthy of respect as if they have no value. The practical part is to treat people and to act in a way that reflects our opinions, it means to truly treat people as less valuable members of society, for example, to discriminate against them. Degrading can be done in two ways: 1) via a face-to-face confrontation where the speaker speaks to the target directly or 2) in absentia where the target is nowhere in sight, but the speaker is addressing

[55] As Fricker (2007) explained.

other hearers. Notice that the degradation of the target happens despite the target not being present at the time of the utterance. The slur need not be said directly *to* the target, it is enough that it is said *about* them. This is an important point because, as we have seen in the first Chapter, when it comes to hate speech, the perlocutionary effect is important, i.e., the spreading of hatred is important.[56] By degrading the target in this sense, the speaker causes *harm* to the target. Since harming is central to this thesis, I will elaborate on this further in point 3) and the later text.

2) Second, the speaker's intention is for their audience to agree with them. They want the hearers to endorse what they are saying about the target and act accordingly (avoid, discriminate against, be disgusted by, etc.). This goal of the speaker applies in case the speaker can be deemed epistemically culpable.
3) Finally, when using a slur the speaker causes *harm*. The harm done by slurs is caused by evoking a negative identity prejudice and by degrading the target as described in point 1). In the section about how stereotypes and prejudice may affect us, we have seen that their effect may indeed be harmful. It is worth noting that harming is independent of the speaker's culpability. Harming happens whether the speaker is aware of the real meaning of the slur or not. The only difference is that, in the case where the speaker is non-culpable, the harm done will be a one-off occurrence and in the case the speaker is aware of his actions, the harm can be systematic which can produce a long-lasting effect on the target. By doing so, the speaker does something ethically and morally wrong, i.e., they are engaged in derogatory-labeling injustice. To portray how exactly harm works and how it produces derogatory-labeling injustice, it is best to turn to the target's perspective.

4.2.3. The target

First, let me note that the targets of slurs that produce derogatory-labeling injustice are members of historically marginalized groups that are at a disadvantaged position in a society just by being members of that group. That's why their already fragile social standing (or something that Waldron (2012) calls dignity) is more susceptible to (further) degrading. It should be a goal of a liberal community to work towards creating a more equal environment for all members, and slurs disrupt the already fragile ecosystem of the community. When it comes to the effect on the target by slurs, the first and most important thing to mention is that

[56] This point was also made in Perhat (2012) and Miščević (2016).

the target is harmed by the slur usage. As mentioned before, the harm done by slurs is caused by evoking a negative identity prejudice and by degrading the target. It is worth noting here that harming happens whether the target is aware of it or not. Since harm is produced by the negative identity prejudice evoked, each time a slur is uttered harming occurs even if the target is not aware of it. This is an important point because this distinguishes harm from mere offense. I have already discussed the difference between harm and offense in the first Chapter where Feinberg understood offense to mean that one is in a mental state of a disliked taste (Feinberg 1985) and where he views offense to be less serious than harm. Waldron (2012) also states that offense is a subjective reaction and mentions that what warrants protection from the state is an attack on a person's dignity. For Feinberg, there are certain harms that can wrong us and that occurs when our interests that are required for our well-being are invaded. After offering an account of how harm may affect the target, I will claim that the sum of what happens to the target and target groups affects the opportunities to acquire primary goods and, in that sense, the harm becomes a wrong which produces derogatory-labeling injustice. But, before examining what kind of harm slurs do, let's remember the difference between harm and offense mentioned in Chapter I.

As stated earlier in Chapter I, offense is a subjective matter and people can be (or not be) offended by various things. Thus, a target may not even feel offended at all by the slur directed at her. That is precisely because offense is a feeling and people feel differently. Let's portray this with an example. My two friends and I are owners of our new company and are having a meeting with a potential investor. My friend A and I are waiting for our friend B who is late. My friend A might feel offended after waiting for 10 minutes and I might feel offended after waiting for 20 minutes or I might not feel offended at all. But, let's say that our friend B's lateness is making us lose money and is damaging our career prospects because the potential investor we are meeting was very clear about valuing his own time and that he only works with companies that respect this. Let us also suppose that my friend A is aware of this and that I am not. Our friend B's behavior is clearly causing both of us harm. The fact that my friend is aware of it and I am not does not, in any way, ameliorate the harm being done to me; I am still losing money, my career is still in jeopardy, my career prospects are in jeopardy and my reputation in business circles is being ruined which in turn means that I may lose potential clients. Moreover, our friend B is not only causing harm directly to us, but he is also harming the reputation of our entire company which affects the people who work for us, he is, so to say, giving our company a bad reputation that can spread to other investors so that our company's business standing is also in jeopardy. The only difference is that, by knowing this information, my friend A might feel anxiety and I, being in the dark, might be spared of this feeling for now. But, in all other aspects, the harm done by our friend B is there, spread through various dimensions. We can translate this example to what happens to the targets of slurs. They too might or might not be aware of the harm being done to them, but,

either way, the harm is there. This harm, done by uttering a slur, can be manifested in the two following effects:

The target can feel an immediate effect of the slur, the *primary effect*[57] of the slur. For this effect of the slur, the target has to be present at the time of the utterance. This effect stems from the immediate flight-or-fight response a slur can have on the target. As Lawrence (1993) explains, hate speech (thus, slurs) can disable the target from any reasonable response, and responses may vary from violence to silencing. In addition to that, the target can also feel degraded and threatened at that very moment which can also produce responses Lawrence described.

The *secondary effect* of the slur has a more long-lasting effect and is more congenial to the consequentialist view because by using hate speech the speaker perpetuates and supports the unjust system of values that some members of the community, mainly the already historically marginalized groups, are inferior to others. For this effect to occur, the target need not be physically present at the time of the utterance. The effect is felt whether the target is present or not and that's because the harm is actually caused by the effect identity prejudice has on the target and for this to happen, just as in the case of stereotypes and prejudice, the effect of harm is independent of the physical presence of the target. Since slurs evoke identity prejudice and since identity prejudice, as Fricker (2007) pinpointed, follows us through every aspect of our social lives, that means that this effect actually echoes through the social dimension and may affect the target in more ways than just in a discourse setting. Since harm is, in addition to degrading, caused by identity prejudice, the next task at hand is to see what kind of empirical evidence[58] could pinpoint the effect of stereotypes and prejudice. In other words, the harms I will list are traced back to stereotypes and prejudice, and since identity prejudices are tied to slurs, these harms, if slurs are used systematically, can be produced by slurs. Since this secondary effect of harm slurs do affects different parts of one's social life, it is best to list these harms as follows:

a) Producing stereotype threat

---

[57] Similar distinction already exists in literature (Leader Maynard and Benesch 2016; Marques 2019, and others), namely the distinction between directly and indirectly harmful speech connected to dangerous speech where the speech is easily disseminated, the speaker is powerful and the audience is likely to condone violence (Leader Maynard and Benesch 2016) and where speech can harm by offense or denigration, where it may thwart democratic processes such as deliberation and so on (Leader Maynard and Benesch 2016). Marques (2019) complements this distinction of dangerous speech and describes "directly harmful speech as language use that is *conventionally* or *constitutively* harmful, using denigrating and derogating language" (Marques 2019, 554). She characterizes "indirectly harmful speech as that which exploits other pragmatic means of communication, like code words, racial figleaves, or meaning perversions" (Marques 2019, 555).

[58] I am well aware that when one refers to empirical research, the research may be inconclusive, i.e., one can find other empirical research that may be a counterargument. However, despite this, I think that we shouldn't disregard empirical findings and that they need to be taken into account.

First, it can produce a stereotype threat where the target's performance is hindered. Stereotype threat was first described by Aronson and Steele (1995). It is mostly tied to the academic setting where targets are known to underperform on tests, but there has also been empirical evidence that stereotype threat can have an effect on one's self-efficacy, reduced ability to pursue certain careers, as well as other physical manifestations such as anxiety.[59] It is superfluous to emphasize how important all of these issues are for an individual. Liberal societies try a great deal to enable equality of opportunities to their members (providing, for example, positive discrimination), so if stereotype threat is something that thwarts these efforts and can even influence one's career choice, then that is something that should be taken seriously.

b) Self-fulfilling prophecy

Second, prejudice can produce a kind of self-fulfilling prophecy, i.e., a person may even internalize the stereotype if it is systematic. Something similar was said by Fricker when she claimed that a person "may be actually caused to resemble the prejudicial stereotype working against her (that's what she comes in some measure to be)" (Fricker 2007, 55). In that sense, one can even suspect that in cases where the harm done by identity prejudice at work in slurs is systematic, this can even affect a person's identity in the sense that the targets may, for example, have lower self-esteem (Swim et al. 2009). As John Donne, an English poet from the 17th century, wrote "no man is an island, entire of itself" meaning that we belong to a community and signifying the importance of society in human development. So, let me reiterate and paraphrase Mead's (1962) quote I mentioned in previous Chapters; according to Mead (1962) we organize our belief system according to the attitudes of our community, we take up our social roles that are expected of us and thus create our personality. So, in other words, the community we are a part of has a great influence on our identity. In that sense, if hate speech is a linguistic device used for communication in a given community and if it is used systematically, we can suspect that this will have an influence on the formation of the identity of the targets. It may shape the way the targets and target groups see themselves and also their role in a community.

c) Maintaining status quo

Third, as already mentioned, the targets are usually historically marginalized groups who are, by being members of such groups, at a disadvantage when it comes to social status.[60]

---

[59] For a more detailed list, please refer to Chapter II.

[60] By social status I mean the status one holds in collective social imagination that follows us in our every social interaction. This status determines how we are treated in social interactions. For example, if we are a

By directing slurs at such groups, the speaker contributes to maintaining the status quo or even further erodes their social status and strengthens their disadvantage. The speaker views the targets as not equal members of the society denying them equal treatment in a community. In the second Chapter, I have listed empirical research on derogatory language and stereotypes that supports these claims. Namely, that derogatory language is linked to the maintenance of status hierarchies (Cervone, Augoustinos, and Maass 2021), that dominant groups tend to use derogatory language more (Rosette et al. 2013), and how derogatory language such as slurs tends to keep social minorities in a subordinate position (Cervone, Augoustinos, and Maass 2021). This point is tied to our interest as members of a community to be actively engaged in our community life which comprises out of, among other things, being able to participate in deliberation. We have seen in the first Chapter how deliberation can be thwarted by hate speech. As Brink (2001) argued, since the speaker disrespects the target by using hate speech, and since the target's reaction to hate speech can disable any reasonable response (Lawrence 1993), it is safe to conclude that hate speech brings no deliberative values to the table, in fact, the effect on the target can be reckoned as harm.[61] Moreover, it seems that exclusion from deliberation is the speaker's goal; the goal is to shun the target and the group the target belongs to because the speaker perceives them as "less than".

d) Hindering deliberation

Fourth, to reiterate the points already made in Chapter I and to add to the point made above about deliberation, Andrew Reid (2019) makes a compelling argument about how hate speech can be detrimental to political discourse in certain cases where there are already injustices and inequalities in place. In such a context, the mutual respect the participants should share could be undermined and the targets of hate speech might be taken less seriously when they decide to participate (Reid 2019). To be taken less seriously means not to fully trust a person, to question their credibility. This is something that Fricker described by testimonial injustice. To strengthen Reid's point, I will add that Fricker (2007) also emphasized how social stereotypes can linger "in our psychology and affect the hearer's pattern of judgment even when our belief system is not in accordance with this" (Perhat 2016, 237). Fricker's example of this is of a feminist not taking her colleagues seriously. In fact, this stealth mode of stereotypes leads Fricker to believe that testimonial injustice happens on a regular basis, and she agrees with Judith Shklar (1990) that "injustice is a normal social baseline" (Perhat 2016, 237). With injustices and inequalities being a normal part of our

part of socially marginalized group targeted by stereotypes and prejudice that target our intellectual abilities, then we can fall victim to stereotype threat.

[61] Stanley Fish makes a similar distinction in his highly influential work There's No Such Thing as Free Speech…and it's a good thing too (1994).

social lives, it is perhaps not a stretch to imagine the negative effects of hate speech on political discourse Reid mentions. In such a social context, hate speech, especially slurs that harbor identity prejudice, may further erode or, at least, sustain said injustices and inequalities. Furthermore, testimonial injustice can contribute to participants of deliberation being taken less seriously and I have previously described how slurs may contribute to testimonial injustice. Quill Kukla's (writing as Rebecca Kukla) notion of changed uptake[62] may also contribute to these points.

> When members of any disadvantaged group face a systematic inability to produce certain kinds of speech acts that they ought, but for their social identity, to be able to produce—and in particular when their attempts result in their actually producing a different kind of speech act that further weakens or problematizes their social position—then we can say they suffer a *discursive injustice.* (Kukla 2014, 441)

So, in Kukla's view, the uptake of a speech act changes in the sense that the speech act type A a person wanted to convey becomes speech act type B due to the speaker being in a socially disadvantaged and disempowered position, for example, due to being a woman. Kukla's point can be best described by an example they provide. Kukla introduces us to the female manager of a factory where almost all other workers are male. Since she is a manager, she uses imperatives to tell the workers what to do, for example, "Your break will be at 1:00 today" (Kukla 2014, 445). But, the workers think of her as a bitch and most of the time they do not comply. As Kukla explains, one possible reason for their failure to comply could be that they are sexist. But, it also may be the case that, instead of taking her speech acts as orders, "because of her gender her workers take her as issuing *requests* instead" (Kukla 2014, 446). The performative force of her speech act is less empowering which, in turn, strengthens her disadvantage. It is easily noticeable how the use of slurs (Kukla mentioned that the manager's workers think of her as a *bitch*, and it is easy to imagine that they also call her that amongst themselves) can strengthen Kukla's point. Being a victim of hate speech can only worsen one's already socially disadvantaged and disempowered position, to borrow from Kukla's terminology. Thus, I feel that the points made by Kukla can also contribute to Reid's notion that hate speech can cause targets to be taken less seriously in deliberation.

e) Impeding opportunities to acquire primary goods

Some of the effects discussed earlier have been mentioned in the literature, although maybe not directly connected to hate speech (such as stereotype threat) but surely mentioned in some contexts. However, to my knowledge, the ability of slurs to impede opportunities to

[62] On Kukla's view "the uptake of a speech act is others' enacted recognition of its impact on social space" (Kukla 2014, 444).

acquire primary goods has not been discussed so far. Therefore I introduce a novel stance on the issue of primary goods.

Each person has specific interests that are important for her to pursue a good life. There are many interests a person can have, but, the two fundamental interests that all others derive from are an interest to be free and to be equal in society which is necessary for persons to pursue a good life according to their understanding. According to Rawls, to be free and equal is a prerequisite to acquiring primary goods, and also each citizen has a fundamental interest in acquiring these primary goods which are the basic rights and liberties, freedom of movement and free choice, the powers of offices, income and wealth, the social bases of self-respect (Rawls 1971). I claim that the harm done to the target by the use of slurs affects the opportunities to acquire these primary goods, i.e., that slurs thwart the target's opportunities to gain them. The effect may be stronger or lesser in some areas. Let me elaborate on this notion further. For example, the effect of the slur on the social bases of self-respect of the target will be greater since the stereotype and prejudice in the slur erode the target's social status, and also, empirical evidence gathered from diary studies by Swim et al. (2009) shows that targets have lower self-esteem.[63] Here, we should note, as Baccarini (2010) pointed out, that Rawls sees self-respect not as an attitude that we have towards oneself, but in terms of institutional facts that support it. This in turn means that it would be possible to establish institutional solutions, as well as the attitude and behavior of individuals that support self-respect (Baccarini 2010). The effect may also be strong in terms of acquiring income and wealth since stereotypes and prejudice spread and sustained by the use of slurs contribute to the marginalization of already marginalized minorities who may have a harder time proving their worth in a society that perceives them as not hard working. In order to acquire income and wealth, the first step is usually to have a good education. However, we have seen that stereotype threat presents a challenge to this since it may have various negative effects on one's academic performance; in fact, studies show that it can even negatively influence one's career choices. Besides stereotype threat, we should also recall from Chapter II how some research found that students' performance can be influenced by the teachers' expectations (Rosenthal and Jacobson 1968; Crano and Mellon 1978; Madon and colleagues 2001). The effect may be lesser in terms of acquiring basic rights since each citizen in a liberal society needs to be equal in the eyes of the law, which is generally the case in liberal societies. But, such guarantees may also be questionable. As we have seen from some empirical evidence in Chapter II, some minority groups may be labeled as more violent due to their group membership and one can suppose that may lead to unfair and unjust treatment in a courtroom. Thus, we can suppose that even basic rights most of us take for granted may be sullied. Freedom of movement, I think, would be less influenced by hate speech since discrimination laws are enforced in liberal communities. Of course, there are many situations that may

[63] Self-esteem and self-respect are different notions, but slurs target both.

influence the opportunities for a person to pursue primary goods, such as being poor or being a member of a minority. But, Rawls requires all should have equality of opportunities, regardless of their background. Also, similarly to how Waldron (2012) borrowed from Rawls when he claimed that in a well-ordered society "everyone can enjoy a certain assurance" (Waldron 2012, 83) that other members of society will act justly, I will do the same and point to a specific requirement of Rawls' Difference principle. The Difference principle states that the distribution of the goods has to benefit all, i.e., the social and economic inequalities where some members have more are just if and only if it benefits the worst off (Rawls 2001). Slurs, as I understand them here, mostly target historically marginalized groups who are already at a disadvantaged position in society and, in order to reach or, at least, come close to a well-ordered society, the government should try to better the positions of the worse off by imploring some political mechanism that would enable these groups to be better off, to some degree. This is not unheard of in liberal societies, in fact, most liberal societies already have certain mechanisms in place, such as positive discrimination. If hate speech via slurs thwarts opportunities to gain primary goods, and I claim that it does, then the state should take certain actions in order to ameliorate the harm this kind of speech does so that it could better the position of those who are worst off. Waldron (2012) has a similar conclusion:

> ...when people call for the sort of assurance to which hate speech laws might make a contribution, they do so not on the controversial *details* of someone's favorite conception of justice, but on some of the fundamentals of justice: that all are equally human, and have the dignity of humanity, that all have an elementary entitlement to justice, and that all deserve protection from the most egregious forms of violence, exclusion, indignity, and subordination. Hate speech or group defamation involves the expressed denial of these fundamentals with respect to some group in society. And it seems to me that if we are imagining a society on the way to becoming well-ordered, we must imagine ways in which these basic assurances are given. (Waldron 2012, 82-83)

Let me now focus on some criticism that was put forth towards Waldron's understanding of dignitary status, or rather the focus he puts on hate speech's influence on it since I feel it would be useful for the discussion at this point. To reiterate, Waldron (2012) claims that each person has dignity which he understands as one's social standing and he thinks the state should protect a person's dignity against hate speech since hate speech is harmful to it. Some authors have suggested that "the reputation of hate speech's potential victims, and the degree to which their equal civic status is recognized by their fellow citizens depends on many factors, of which the circulation of hate speech is just one" (Seglow 2016,

1108).[64] This is true, and I see two possible replies. First, I don't see it as problematic that one's social standing may be influenced by various mechanisms and only one of them being hate speech. As I have argued above, liberal societies already have certain mechanisms in place to ameliorate other disadvantages people may face, one of them being positive discrimination laws for minorities. So, hate speech regulatory laws would be just one more mechanism in place to secure a better outcome for people from vulnerable groups who are usually targeted by hate speech. Second, considering the focus on hate speech and hate speech laws, as Shiffrin (2014) pointed out, communication is key if we want to be known by others as individuals we are and speech is crucial for this. Moreover, I have previously recalled instances where speech was critical in determining one's life, for example, the case of naming one a witch, the case of Galileo Galilei, and the case of Jews in Nazi Germany. In all of these instances, speech played a crucial part. So, in that sense, we can consider speech to play one of the most important roles in society and, consequently, focusing on the harms speech can do is not undesirable. Considering these criticisms of Waldron's account, some authors have taken a slightly modified approach. Namely, Seglow (2016) has claimed "that hate speech directly undermines the self-respect of hate speech's victims and does not serve the self-respect of hate speakers" (Seglow 2016, 1115). For Seglow the focus should be less on dignity and more on the impact hate speech may have on self-respect which he considers to be very important in one's perception of oneself. Hate speech can undermine self-respect in three ways: by thwarting an interest vulnerable citizens may have in deliberation on their aims, by implying, through hate speech, that the target's thoughts are not as important thus weakening the target's belief in her aims, and by weakening deliberation since hate speakers consider that their targets have nothing worth contributing (Seglow 2016). Both Seglow's and Waldron's views have merit, and, if we agree that hate speech is fueled by stereotypes and prejudice, namely identity prejudice, then it is easy to see how stereotypes and prejudices can do harm in the way described both by Waldron and Seglow. In fact, I think that Seglow's account of self-respect could be added as an extra argument to the claim that hate speech may thwart one's opportunities to gain primary goods, one of them being the social basis of self-respect.

f) Hindering thinkers' interests

Finally, to conclude what constitutes the secondary effect of harm on the target, it is useful at this moment to look back at what was said in the first Chapter, in the discussion about freedom of speech. Namely, I think I can now provide an answer to Seana Shiffrin's account of the thinker-based approach to freedom of speech. To my knowledge, no such answer has been given in the literature that would provide a satisfactory account of why hate

[64] See also Robert Simpson 2013.

speech actually hinders thinkers' interests. Shiffrin (2014) holds that communication is key to conveying to others our thoughts so that they can get to know us as individuals we are. To accomplish that there are certain interests we have as individual thinkers. For Shiffrin, these interests can be fulfilled only by communicating freely and getting feedback on our thoughts from others, i.e., free speech is crucial. But, after listing the potential harms that hate speech does, my intuition is that hate speech can thwart these interests and that unregulated free speech is not the way we can achieve them. I will therefore list Shiffrin's interests here (again) and try to explain why the harm hate speech does could be detrimental to them:

> a. *A developed capacity for practical and theoretical thought*. Each thinker has a fundamental interest in developing her mental capacities to be receptive of, appreciative of, and responsive to reasons and facts in practical and theoretical thought, i.e., to be aware of and appropriately responsive to the true, the false, and the unknown.
>
> b. *Apprehending the truth*. Each thinker has a fundamental interest in believing and understanding true things about herself, including the contents of her mind, and the features and forces of the environment from which she emerges and in which she interacts.
>
> c. *Exercising the imagination*. In addition, each thinker has a fundamental interest in understanding and intellectually exploring non existent possible and impossible environments. Such mental activities allow agents the ability to conceive of the future and what could be as well as what could have been. Further, the ability to explore the nonexistent and impossible provides an opportunity for the exercise of the philosophical capacities and the other parts of the imagination.
>
> d. *Moral agency*. Each thinker has a fundamental interest in acquiring the relevant knowledge base and character traits as well as forming the relevant thoughts and intentions to comply with the requirements of morality. (This interest, of course, may already be contained in the previously articulated interests in developing the capacity for practical and theoretical thought, apprehending the truth, and exercising the imagination [a-c].)
>
> e. *Becoming a distinctive individual*. Each thinker has a fundamental interest in developing a personality and engaging more broadly in a mental life that, while responsive to reasons and facts, is distinguished from others' personalities by individuating features, emotions, reactions, traits, thoughts, and experiences that contribute to a distinctive perspective that embodies and represents each individual's separateness as a person.

> f. *Responding authentically*. Each thinker has a fundamental interest in pursuing (a-e) through processes that represent free and authentic forms of internal creation and recognition. By this, I mean roughly that agents have an interest in forming thoughts, beliefs, practical judgments, intentions, and other mental contents on the basis of reasons, perceptions, and reactions through processes that, in the main and over the long term, are independent of distortive influences. In saying these processes are independent of distortive influences, I mean that the choices of what to think about and the contents of one's thoughts do not follow a trajectory fully or largely scripted by forces external to the person that are distinct from the reasons and other features of the world to which she is responding *qua* thinker. So, too, thinkers have an interest in revealing, sharing, and considering these mental contents largely at their discretion, at the time at which those contents seem to them correct, apt, or representative of themselves, as well to those to whom (and at that time) such revelations and the relationship they forge seem appropriate or desirable. These are the intellectual aspects of being an autonomous agent.
>
> g. *Living among others*. Each thinker has a fundamental interest in living among other social, autonomous agents who have the opportunities to develop their capacities in like ways. Satisfaction of this interest does not merely serve natural desires for companionship but also crucially enables other interests *qua* thinker to be achieved, including the development of self and character, the acquisition and confirmation of knowledge, and the development and exercise of moral agency.
>
> h. *Appropriate recognition and treatment*. Each thinker has a fundamental interest in being recognized by other agents for the person she is and having others treat her morally well. (Shiffrin 2014, 86-88)

It seems to me that hate speech could thwart reaching some of those interests. For example, in order to develop authenticity, one would need to be able to form mental content independent of distortive influences, as Shiffrin writes. If we follow Moles' (2007) argument, then mental contamination would not allow us to be truly autonomous. As Moles notes, some external forces, namely society which is riddled with stereotypes and prejudice, may lead to us having responses and unconscious processes with which we cannot identify. So, for example, a person who is not racist may have some racist responses despite their belief system not being in accordance with racist ideology. Similar thoughts can be found in Fricker (2007) when she claims that stereotypes can have an effect on us and our judgment without us being aware of it. In Chapter III, we have seen that there is empirical research that can corroborate these claims about unconscious processes (Devine 1989; Correll and colleagues

2002). Thus, it seems that stereotypes and prejudice may present distortive influences that disable us from forming mental content that is truly independent. This could also be tied to the issue of forming one's identity, or, as Shiffrin notes, becoming a distinctive individual. I have, in the previous text in Chapter II as well as this Chapter, already argued that stereotypes and prejudice may have an influence on our identity in the sense that they can serve as a self-fulfilling prophecy if the target internalizes the stereotypes (Fricker 2007), or that targets can have lower self-esteem (Swim et al. 2009). Since we live in a community, and as Shiffrin also lists living among others as one of the interests of persons qua thinkers, we should take into account that the community we are a part of influences our belief system and thus influences our personality (Mead 1962). I have argued that slurs, if used systematically, may negatively influence the way targets see themselves and their role in a community. Of course, one can argue that it is obvious that the society we are emerged in has an influence on our identity, which is true, as we have seen in Chapter II, in the discussion about socialization, but what I see as problematic is when what we take from society are prejudices about other minority members, or, if we are targets of hate speech, the potential internalized belief that we are "less than". One can be a distinctive individual even with all those beliefs, but, I am left wondering what it would look like if we were able to achieve our full potential, without being negatively influenced by hate speech, be it as hearers or targets.[65] Shiffrin also emphasizes the need to be able to apprehend the truth and to be responsive to the true, the false, and the unknown. However, stereotypes and prejudice, especially identity prejudice, tend to create an environment where there is typically resistance to any counterevidence, so perhaps it is the case that this goal would be reached more easily in an environment where there are fewer stereotypes and prejudice. Finally, for appropriate recognition and treatment where the thinker is recognized for the person she is and where others treat her morally well (Shiffrin 2014), hate speech presents an obvious challenge. In the text in this Chapter, I have argued that hate speech degrades the social status of already marginalized groups and that, by using slurs, the speaker views them as not equal members of society. In fact, derogatory language, such as slurs, tends to keep social minorities in a subordinate position (Cervone, Augoustinos, and Maass 2021). Since hate speech harms its targets in the ways described in this Chapter, I think that Shiffrin's account would benefit more from regulating hate speech than the other way around.

4.2.4. The listener

[65] One can argue that hate speech isn't the only negative influence we face in our social lives, and I have already provided possible answers in this Chapter.

Listeners are not only passive bystanders. Their role is important in battling hate speech, but we will review that specific role in the chapter to come. For now, I would like to focus on the effect slurs may have on listeners.

There is empirical evidence (Devine 1989; Correll and colleagues 2002) that suggests that there is some automatic stereotyping and automatic responses[66] due to stereotypes and prejudice that, as Fricker says, reside in the collective imagination. Fricker (2007) was aware of this stealthy notion of stereotypes and prejudice and illustrated it with an example of a feminist who, unfortunately, takes the word of her female colleagues less seriously. Moles (2007) calls these automatic responses mental contamination and views them as a possible threat to one's autonomy if our belief system is not in accordance with social stereotypes and prejudice. This would potentially mean that listeners are in danger of having responses they, in good conscience, could not abide by—which is problematic.

Cervone, Augoustinos, and Maass (2021), in their paper about the consequences of derogatory language and hate speech, have provided an encompassing portrayal of empirical research that considers the effects on listeners, as follows:

> …bystanders who are more exposed to hate speech become desensitized to it and consequently perceive it as less offensive and more acceptable (Soral et al., 2018; Winiewski et al., 2017). This may lead them to not recognize how the derogatory language they are exposed to affects their own attitudes and behaviors, such as nonverbal cues (Goodman et al., 2008), charity giving (Ford et al., 2008) and radical political attitudes and behaviors (Soral et al., 2018; Winiewski et al., 2017). Exposure to derogatory language against a specific minority group also leads to greater distancing, both physical and social, from its members. In one study, after being subliminally exposed to homophobic epithets, heterosexual individuals tended to sit further away from a gay man they expected to meet (Fasoli et al., 2016). Furthermore, people more frequently exposed to hate speech toward a certain community are also less willing to have social contact with its members (Winiewski et al., 2017; see also Soral et al., 2018). (Cervone, Augoustinos, and Maass 2021, 89)

They also note how:

> ...derogatory language may affect societies beyond discrimination: people (especially youth) more exposed to hate speech deem other non-normative behaviors as more socially and morally acceptable, as well as worth imitating (Winiewski et al., 2017). Thus, according to Winiewski et al. (2017), derogatory language may elicit an effect similar to that described

[66] For reference to the research, please refer back to Chapter II, section on Socialization.

> by the *broken windows theory*, that is, that small indicators of disorder encourage anti-social behaviors and crime by signaling that those behaviors are the norm (see Welsh et al., 2015). (Cervone, Augoustinos, and Maass 2021, 89-90)

Listeners are already exposed to stereotypes and prejudice that preside in the social imagination which means that they bring their own bias into the discourse. Additionally, they can also have some automatic responses, as described earlier.

Let's remember that the speaker wants the audience to agree with them, and sometimes the audience can feel great pressure to do so. For example, in cases when the speaker is in a position of power (the example of an executive of a company who has identity power (because of being a man) and economic power (because of being an executive)), the audience may even assign such speakers with credibility excess, in which case the pressure to agree with them is even greater.

Because of all that was mentioned above, listeners have to be epistemically responsible agents in order to be able to recognize stereotypes and prejudice, not just the speaker's, but also their own.

After reviewing harm from three perspectives, the speaker's, the target's, and the listener's, we are finally able to offer an account of derogatory-labeling injustice.

4.3. Derogatory-labeling injustice

Finally, after reviewing the harm that is inflicted on the targets, we can say that by inflicting harm in this way, the target is actually wronged, i.e., the speaker is engaged in something we might refer to as derogatory-labeling injustice (inspired by Fricker's and Kukla's notions). Let me reiterate a few points here in order to place derogatory-labeling injustice in the social context where other injustices (such as testimonial and discursive) occur.

First, identity prejudices follow us through every social dimension, and they stay with us in a discourse setting as well. Slurs are fueled by identity prejudice and each time a slur is used, identity prejudice is utilized. According to Fricker:

> systematic testimonial injustices, then, are produced not by prejudice *simpliciter*, but specifically by those prejudices that 'track' the subject through different dimensions of social activity—economic, educational, professional, sexual, legal, political, religious, and so on. Being subject to a tracker prejudice renders one susceptible not only to testimonial injustice but to a gamut of different injustices, that and so when such a prejudice generates a testimonial injustice, that injustice is systematically connected with other kinds of actual or potential injustice. (Fricker 2007, 27)

This tracker prejudice Fricker talks about is identity prejudice, the one we explained is related to our social identity. Even though Fricker allows for identity prejudice to be positive or negative, she, as well as I, is concerned only with negative identity prejudice.[67] Testimonial injustice is fueled by negative identity prejudice in the hearer which then distorts their judgment about the speaker resulting in the speaker receiving a credibility deficit and they fail to pass on their knowledge. The ability to convey knowledge is, for Fricker, a central interest of us as humans. In the same fashion, slurs are fueled by negative identity prejudice and when a slur is used, negative identity prejudice is utilized. With the systematic use of slurs, derogatory-labeling injustice is born. In the previous text, I have exhausted a number of harms that stem precisely from prejudice and that harm something we may consider to be our important interests. As Fricker notes, negative identity prejudice may produce not only testimonial, but other injustices as well, and, in the previous passage, I have utilized empirical evidence to show how prejudices that stem from the use of slurs, if used systematically, thwart a person's important interests. These interests are: a) being successful in one's academic life which is a foundation for economic stability later in life and career

[67] Except in the case of appropriated uses of slurs where the negative identity prejudice is replaced with positive identity prejudice.

opportunities (which can, as we have seen, also be influenced by stereotype threat); b) being able to form one's identity free of the kind of influences due to which we can form a negative image of ourselves or our role in a community; c) being able to hold a social standing in a community that will enable us to have equality of opportunity; d) being able to participate in deliberation equally; e) being able to acquire primary goods; f) being able to pursue interests that will enable us to communicate our thoughts to others so that others may come to know us as an individual that we are. All of the mentioned interests may be obstructed by the systematic use of identity prejudice directed at targets via slurs. Such harms produce a novel kind of injustice, an injustice I refer to as derogatory-labeling injustice.

Second, in order to understand the interplay between hate speech, slurs, and derogatory-labeling injustice, we need to establish domains for each category.

At the moment of writing this thesis, there still seems to be a lot of confusion in the literature as to what constitutes hate speech. As we have seen, hate speech has been a long-debated subject and no unified definition has been agreed upon. Even though there is no unified definition of hate speech, when examining existing various definitions and documents, there can be found some characteristics that fit the label: it is usually considered to be public speech targeted at an individual or a group of people with ascribed characteristics such as race, gender, ethnicity, and so on. Hate speech can take many forms, such as symbols (swastika) or a deed (cross burning), but the most used way of expressing hate speech is words. One of the most used vehicles of hate speech is slurs, as I have previously emphasized. However, not all slurs will fit into the category of hate speech. For example, we can imagine a group of friends who just had a falling out where one of them calls the other a *bitch*. That would hardly be considered hate speech since certain conditions for this utterance to fall into that category haven't been met. All of this will, of course, also depend on the view one takes on hate speech and how one understands it. That means that, depending on the position we take on hate speech and what constitutes hate speech, we will have a different understanding of which slurs fall into the category of hate speech. But, generally, to reiterate what was previously said: some slurs will be considered hate speech, and some will not. Let's remember the diagram that portrays this:

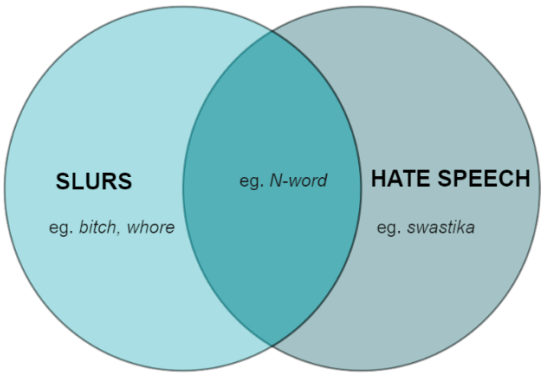

Figure 1

As I previously emphasized, I will not be advocating a new definition of hate speech here. My aim here is to pinpoint that, in certain cases, some slurs may produce derogatory-labeling injustice. As Jeshion explained slurs "signal that their targets are unworthy of equal standing or full respect as persons, that they are inferior as persons" (Jeshion 2013a, 308). The literature on slurs has, up to the point of writing this thesis, been concerned with explaining the semantics and pragmatics of slurs, as well as their ethical implications. Fricker (2007) presupposed that there are other injustices at play in society. I think that slurs, if certain conditions are met, may produce derogatory-labeling injustice. For slurs to produce derogatory-labeling injustice, they have to be used in their literal sense to degrade, they have to be directed at historically marginalized groups, and they have to be used by someone in a position of power. For example, the Croatian slur *Tovar* is a slur but not for a marginalized group—it refers to a group of football fans from the town of Split. Another example would be referring to political opponents by derogatory terms such as *libtards*. Or, the Croatian phrase *uhljeb,* which is used to refer to people who take up various job positions in civil service and are thought to have been employed in these positions due to nepotism. By historically marginalized groups I understand any groups with certain ascribed characteristics that have a history where they have been marginalized in society in terms of not having their rights respected (such as women, gays, people of color, etc.). Usually, this applies to minorities. However, when talking about minorities one needs to be careful not to refer to quantity because there were cases, such as the apartheid in South Africa, where the white minority oppressed the black majority. This issue can be portrayed by the diagram as follows:

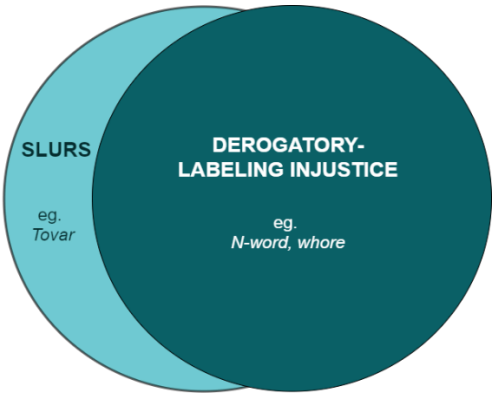

Figure 3

On the one hand, some slurs will produce derogatory-labeling injustice if certain conditions are met. On the other hand, there are slurs that will not produce derogatory-labeling injustice because they do not meet the requirements.

In addition, for derogatory-labeling injustice to occur, slurs have to be used systematically. When slurs are used systematically, each instance of using a slur (in the literal sense to degrade) accumulates and can, over time, create this novel form of injustice.[68] Fumagalli (2021) also considers this cumulative harm: "…we might read public hate-speech events as contributions to a public hateful environment. Harm would be a long-term cumulative harm that accretes discriminatory attitudes and behaviors" (Fumagalli 2021, 12). Even though when Fumagalli (2021) talks about cumulative harms, he means the harm produced by public hate speech, I will add that slurs that are not considered hate speech (if we define hate speech as public speech) may as well produce derogatory-labeling injustice and the harms I described in this thesis, precisely because of what slurs imply. Another thing worth emphasizing at this point is that slurs, even when they are directed at an individual, also target a group. For example, when one calls Anna a *whore*, they are also saying something about all women. To reiterate from *Table 2* in this thesis, the word *whore* refers to a woman who sleeps with a lot of men and is therefore bad. The underlying negative identity prejudice is that women, in general, should not sleep with a lot of men because that is not lady-like, and being lady-like is something all women should aspire to be. If they go against this prejudice, that is considered to be bad. The same applies to other slurs, such as calling John the N-word, in which case we not only say something bad about John but also evoke a negative identity prejudice about people of color in general. This means that with every slur usage, we refer not only to the individual but to a group of people that individual belongs to.

[68] Similar argument was developed by Alexander Tsesis in his book Destructive Messages (2002).

Furthermore, the harms mentioned and defined in the previous text are produced by, but not limited to, slurs. That means that these harms could be produced by other means and not just slurs or hate speech. However, not every case of producing these harms would be derogatory-labeling injustice. To reiterate, derogatory-labeling injustice is produced only when certain conditions are met. As previously described, it happens in a discourse setting where specific utterances have to be produced, either face-to-face or in absentia. By specific utterances I mean slurs used for degrading and which target specifically historically marginalized groups. And, secondly, the speaker has to be in a position of power. Let me elaborate on this last point a bit further which brings us to the third clarificatory point.

Namely, for derogatory-labeling injustice to occur, it has to be enacted by those who hold certain social power. For example, it can be enacted by a man over a woman where the man has identity power over a woman, or it can be enacted by a senior executive over his female colleagues in which case he has both the hierarchical and identity power over them. As described earlier, these various powers mean that the listeners might find themselves under certain pressure to agree with the speaker. This pressure need not be evident, and listeners may even not be aware of it since the pressure stems from a systematic power imbalance that exists in all social contexts.[69]

Thus, I am finally in a position to offer a definition of derogatory-labeling injustice. We can say that *derogatory-labeling injustice happens in a discourse setting where the speaker, who is in a position of power, by using derogatory language, i.e., slurs, labels the target with negative identity prejudice, and thus wrongs the target by harming one or more of their important interests*. In other words, when the harms mentioned above are produced by the (systematic) use of slurs, derogatory-labeling injustice is born. This further degrades the target's social status in a way that thwarts opportunities for the target to pursue primary goods and where the harm is manifested through the immediate threat the target feels, and through a more long-lasting effect of harm the target may endure. Derogatory-labeling injustice targets historically marginalized groups that share some ascribed characteristics and it supports and perpetuates the unjust system of values in the society, i.e., that there are groups of people who are "less than", in other words, by treating them as not equal members of society we take away from their humanity which can produce and /or support and perpetuate this unjust treatment in society. Since slurs target not only individuals but groups, when uttering a slur the speaker not only harms the individual but a whole group because she says something bad about the inherent characteristic of that group. Thus, by uttering slurs, the speaker harms the social status of an entire group and supports and perpetuates prejudice

[69] Let me borrow from Fricker again and agree that there is social power that operates in society which she defines as follows: "a practically socially situated capacity to control others' actions, where this capacity may be exercised (actively or passively) by particular social agents, or alternatively, it may operate purely structurally" (Fricker 2007, 13).

about that group further. This further strengthens the prejudice that already exists in the collective social imagination about certain groups and it keeps them at a disadvantaged position in society. Understood in this sense, derogatory-labeling injustice may pose a threat to some democratic processes, such as deliberation since target groups are discouraged to participate, or, when they do participate, they may be taken less seriously than they otherwise would have.

As said, derogatory-labeling injustice would target already marginalized groups in society. This in turn means that, in a general sense, dominant groups would not be susceptible to derogatory-labeling injustice. Since slurs harbor identity prejudice and identity prejudice is, in many cases (but not always), concerned with historically marginalized groups, derogatory-labeling injustice would also be concerned, by extension, with these groups. These limitations potentially solve the problem of overreaching which has been a real problem for strategies aimed at legal regulation of hate speech. As Kulenović (2023) notes: "overreaching can result in legal bans on hateful speech being used to stifle genuine democratic discourse and sanction legitimate public criticism" (Kulenović 2023, 522). For democratic processes, it is important that citizens are able to criticize those in power, for example, those who hold political power. So, groups such as politicians would not be susceptible to derogatory-labeling injustice, and thus would not warrant protection by regulatory laws. It would be very questionable to restrict any kind of speech when it is used to criticize those in power. In this sense, the speech serves as a kind of corrective measure that every liberal democracy needs. One can of course imagine that there is a possibility of an overlap. For example, we can imagine someone referring to Barack Obama by using the N-word, but in that case, the focus is not on Obama as a politician but as a member of a minority and the word used says nothing about him as a politician but as a person of color, so in that sense he would be a target of derogatory-labeling injustice.

Derogatory-labeling injustice permits us to define and understand what happens to the targets of slurs if the use of slurs is systematic. Up to this point, there have been discussions in literature about the harm derogatory language may inflict, but it has not been systematized. Authors have tried to answer why slurs offend and/or harm but it seems that it was a challenge to provide an elaborate explanation. Although some authors argued that there indeed is a stereotype in the semantics or pragmatics of slurs, negative identity prejudice (described by Fricker) evoked by slurs provides an explanation as to why slurs have a negative evaluative layer and why they cause harm. Evoking negative identity prejudice when a slur is uttered causes harm and gives rise to an injustice foreseen by Fricker, an injustice I described as derogatory-labeling injustice. Since neither the notion of slurs nor hate speech seem to precisely capture and explain what happens when such language is uttered and the exact harm that may be inflicted (and in what way), the notion of derogatory-labeling injustice fills this explanatory gap. Even though the boundaries between slurs, hate

speech, and derogatory-labeling injustice are often blurry, derogatory-labeling injustice serves as a link that explains how derogatory language may lead to injustice. It is a notion that provides us with a clearer way of limiting hate speech in cases where slurs fall into the category of hate speech. On the other hand, in cases where slurs do not fall into the category of hate speech, but may still cause derogatory-labeling injustice if they are used systematically, we have other ways of dealing with this that exclude legal prohibitions. Of course, limiting any kind of speech needs to be carefully addressed, which is something I will say more about in the next Chapter.

For derogatory-labeling injustice to happen, a slur may or may not fall into the category of hate speech. Whether a slur falls into the category of hate speech or not is not a key element for derogatory-labeling injustice to occur. Derogatory-labeling injustice is marked by systematicity, meaning that it may accumulate over time. If a friend, who is in a position of power (male), talking to his other friends privately uses a slur to target a historically marginalized group, his utterance echoes into the collective social imagination. These many similar echoes accumulate over time and may produce derogatory-labeling injustice. Even though sometimes derogatory-labeling injustice can be more immediate, such as in cases where slurs are clearly used as hate speech by someone in power, such as a politician who refers to "these faggots", derogatory-labeling injustice can also be produced by slurs that, in some views, wouldn't be considered hate speech (such as in the example with friends conversing with each other). In cases where derogatory-labeling injustice is produced by slurs that fall into the category of hate speech, we would have a strong case to limit this kind of speech. Finally, we can portray these concepts as follows:

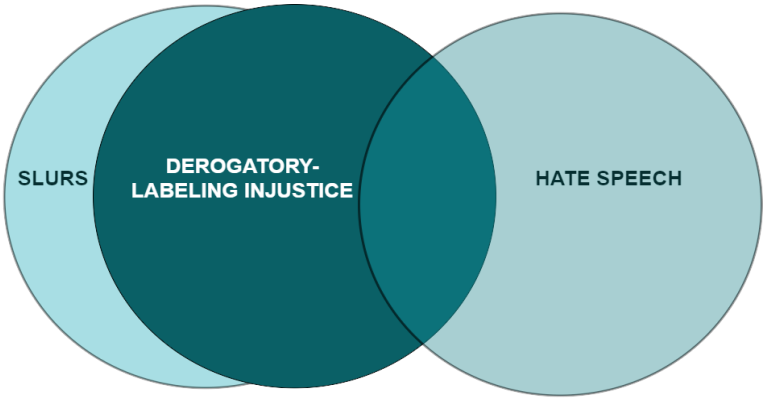

Figure 2

Hate speech as a broader concept may or may not be expressed by words. In some cases, it will be expressed by using slurs as a vehicle of hate speech. Some slurs will thus be hate speech, such as the N-word. These slurs would also produce derogatory-labeling

injustice. However, in cases where the slur would traditionally not be considered hate speech, it can still produce derogatory-labeling injustice, and prime examples of this are gendered slurs for women (such as the slur *whore*). Finally, there are slurs that are not hate speech and that wouldn't produce derogatory-labeling injustice because they don't meet the requirements to do so, i.e., they do not target historically marginalized groups or they are not uttered by someone who has social power (such as the Croatian word *Tovar*).

Let me now portray this with examples. Let's imagine a male influencer with five million followers who is doing a podcast where he targets a certain female celebrity and calls her a *whore*. First, let's consider that by calling her a *whore*, he is also targeting and saying something bad about an entire group of women who are deemed promiscuous. We can also recall, from Chapter III, that I understand the term promiscuous to be broad enough to include, not just having a lot of sexual partners, but *acting* promiscuously as well, which can include a myriad of behaviors such as merely dressing up in a more revealing fashion. So, by calling her a *whore*, the influencer negatively characterized an entire group of women. Second, let's consider the power relations. The influencer holds identity power just by being a male, and he holds what I will call an *influential* power since he may influence the opinions of quite a number of people. The potential to disseminate his views is huge. This kind of slur usage would be able to produce derogatory-labeling injustice. All accounts are satisfied: the slur is used derogatorily, it targets a marginalized group (women), it is used by someone who holds an identity power (being a man), and an influential power.

On the other hand, we could imagine a different kind of situation I previously mentioned between a close group of friends conversing with each other where the potential to disseminate derogatory language to a greater number of people is diminished. In this case, when one of the friends uses the term *whores* to refer to their female friends, this would not be considered hate speech, but it would echo into the collective social imagination accumulating to create derogatory-labeling injustice.

Finally, using the N-word would in most cases be considered hate speech by most courts, if said in a public space. It would clearly also produce derogatory-labeling injustice.

To reiterate, my claim was that the harm inflicted by slurs, categorized into primary and secondary effects, culminates in a novel notion of derogatory-labeling injustice. By shedding light on the nuanced relationship between slurs, prejudice, and harm, this thesis calls attention to the pressing need for understanding and addressing derogatory-labeling injustice in order to foster a more inclusive and equitable society. One of the ways that can be done is surely to examine the possible responses to this phenomenon.

I would like to add that I am well aware that stereotypes and prejudice are not only created by slurs or hate speech and that the effects of stereotypes listed may come from other parts of our social lives, in fact, as Fricker noted, identity prejudice follows us through every social dimension. But, slurs, with identity prejudice being embodied in them, can spread and

further the injustices already at play in society. This is important because almost all documents about hate speech mention the perlocutionary force, i.e. the spreading of hatred. Also, as I have already argued elsewhere, language makes up a great part of our social lives, people rely on communication for a number of reasons; to get their meaning across, for other people to get to know them as the person that they truly are (Shiffrin), to gain knowledge (Fricker), and so on. Language is an essential part of community life and it should be treated that way. So, if there is something in language that creates and enacts harms, it should not be taken lightly. Even though some harms may not be completely remedied just by restricting hate speech, mindful restriction can help to get closer to the ideal, or to the well-ordered society, as Waldron suggested, inspired by Rawls. After all, there are many forms of sanctions. The fifth and final Chapter will be dealing with ways to address these issues. The idea is that we have a strong case for restricting speech that falls into the category of hate speech and that produces derogatory-labeling injustice. Speech that causes derogatory-labeling injustice but isn't considered hate speech also needs to be addressed, but there are other options at our disposal to do so without legal bans.

CHAPTER V: POSSIBLE RESPONSES

5.1. Introduction

In the previous Chapters, I have presented the debate about freedom of speech and hate speech where the main concern for scholars is how to reconcile the protection of the fundamental principle that is freedom of speech and limiting speech that can potentially be harmful. Since there is no consensus about what hate speech is, I have given a prime example from European and American legal systems to show how that may be confusing for regulatory laws since there can be stark differences in the legal treatment of hate speech. I have given my contribution to the debate by focusing on one of the most used hate speech devices—slurs. I have also provided my contribution to the debate by claiming that there is a specific kind of stereotype evoked by a slur: the identity-prejudicial stereotype explained by Fricker (2007). I have then proceeded to elaborate on what kind of harm can stem from this kind of prejudice from the speaker's, the target's, and the listener's perspective. I have based my own research on empirical research on stereotypes and prejudice and their potential effects. Finally, I have concluded that such harm caused by identity prejudice evoked by the slur can be seen as derogatory-labeling injustice. To reiterate, derogatory-labeling injustice happens in a discourse setting where the target is harmed by utilizing the identity prejudice evoked by slurs. I claim there is a pressing need for understanding and addressing derogatory-labeling injustice in order to foster a more inclusive and equitable society. One of the ways that can be done is to examine the possible responses to this phenomenon.

There have been various attempts at providing the most optimal solution that would somehow protect freedom of speech but that would also be effective against hate speech and its potential harms. Usually, there are two camps in the debate about possible solutions. On the one side, there are authors who advocate for legal sanctions of such speech, and on the other are authors who opt for the *more speech* method. Having defined derogatory-labeling injustice, and having explored the boundaries between the triad that is derogatory-labeling injustice, slurs, and hate speech, the possible solution could be found in the middle. More precisely, in cases of slurs that are hate speech that produce derogatory-labeling injustice, we can say we have strong reasons to restrict such speech by legal means. On the other hand, in cases where some forms of speech (slurs) produce derogatory-labeling injustice but are not considered hate speech, we can utilize counterspeech, which can take many forms (such as a moral reprimand). However, counterspeech can be utilized even in cases when derogatory language succumbs to legal sanctions, i.e., one does not exclude the other. In any case, the solution to these issues that plague our society and our language is not simple. I think that providing a solution to how to treat derogatory-labeling injustice and hate speech needs a

complex approach and, in addition, it needs coordination of several strategies, and the involvement of all the stakeholders in society. In other words, in order to be effective, the responses to such phenomena have to be two-fold, i.e., they have to come from two directions. Thus, I will divide the possible responses to slurs and derogatory-labeling injustice into two broad areas:

a) *The first set of responses* is concerned with the responses that can be given by members of society on an individual level and we can center those responses on the perspectives previously established, namely the speaker's, the target's, and the listener's perspective. The said responses can be made individually or as a group. These responses will be concerned with epistemic responsibility, the responsibility we share as epistemic agents active in community life. In that sense, we put the burden on members of society to counter dangerous language. These kinds of responses would fall under the category of counterspeech.

b) *The second set of responses* is concerned with institutionalized responses where the burden to react to problematic language is lifted from individual members and groups in society and placed on the institutions. One can argue that placing such a burden and responsibility on the shoulders of individuals and groups in society is too much to ask. I partly agree with this, and therefore I will advocate for the state's response as well. I will argue that the state has the responsibility to react to harms done to its most vulnerable members by providing certain institutionalized protections. In addition to that, the state should provide resources that can help individuals develop certain epistemic virtues that can, in the long run, help them recognize stereotypes and prejudice at play in society. These kinds of responses can come in the form of legal sanctions when needed (in cases where derogatory language and hate speech overlap) and in the form of counterspeech, this time not given by individuals but by institutions.

The individual and the institutionalized responses go hand in hand and complement each other. I find this symbiosis to be the most effective way to combat injustices produced by prejudice.

5.2. Counterspeech

Before moving on to examining the first and second sets of responses, I feel it is important at this point to say something more about counterspeech since it is an integral part of both the individual and institutional responses. Since counterspeech has gained much attention from authors working on free speech and hate speech and has become one of the most favored ways of responding to problematic speech, I will dedicate this subsection to some of the main questions that arise from it. Literature on counterspeech is in abundance and due to the scope of this thesis, it is impossible to cover all of the issues surrounding this notion, so I will focus mainly on the most prominent ones. That being said, I will not engage in trying to answer some of the posited questions, but merely point them out.

The initial idea of counterspeech came from a 1927 case Whitney v. California where judge Brandeis said that "the solution to bad speech is more speech" (Brandeis 1927). Since free speech is a fundamental right in liberal democracies, restricting it is not taken lightly and many authors agree that it should be reserved just for the most serious cases of hate speech.[70] More recently, in the United Nations' *Strategy and Plan of Action on Hate Speech* issued in 2019, the focus is also on more speech, not less, to battle hate speech (United Nations 2019).

Therefore, counterspeech is focused on counteracting the harm done by other speech (Cepollaro, Lepoutre, and Simpson 2022). As seen in the previous subsection and as Cepollaro, Lepoutre, and Simpson (2022) have stressed, counterspeech can be performed by various agents. It may be performed by targets, by listeners, or by institutions. In addition, "counterspeech's audience is also variable, and this matters because it has an effect on its authority and efficacy. The audience may include the person whose speech is being challenged, victims, bystanders, or any combination of these" (Cepollaro, Lepoutre, and Simpson 2022, 4).

Cepollaro, Lepoutre, and Simpson (2022) have identified two crucial questions that surround the notion of counterspeech and that should, to their understanding, be of interest to philosophers seeking to explore the issues surrounding counterspeech. These are the Efficacy question and the Deontic question. The former is a question of whether counterspeech is an effective way to resist hate speech, and the latter is a question concerning the duties to engage in counterspeech.

Out of the two outlined questions, the main one, according to Cepollaro, Lepoutre, and Simpson (2022), is the Efficacy question. They explain that if counterspeech is ineffective in countering hate speech, then engaging in it makes no sense, but if it is equally

[70] What constitutes most serious cases, as we have seen in Chapter I, depends on our understranding of hate speech. So, in the US those cases would be incitement to violence, whereas European laws are less forgiving when it comes to potential hate speech.

or even more effective than restrictions, then restricting speech should be abandoned since preserving freedom of speech should be the norm. Cepollaro, Lepoutre, and Simpson (2022) give an overview of some of the issues counterspeech may face and which may render it ineffective. As they note, there have been some worries that counterspeech would increase the salience of speech it is supposed to mitigate. The worry about salience is actually the worry that counterspeech may backfire. As some authors have argued, it may be the case that responding to hateful speech may give it unwanted attention and therefore render it credible for others (Levy 2019). Similarly, if an idea becomes familiar some authors fear it may be believed, as well (Lewandowsky et al. 2012). Langton (2018), on the other hand, sees the lack of authority as a possible problem for counterspeakers.

The possible solution to these problems Cepollaro, Lepoutre, and Simpson (2022) have listed, "in developing a good framework for assessing counterspeech's efficacy, is to establish when – in which cases, in connection with which negative outcomes – the efficacy of counterspeech should be judged in terms of epistemic results, and when in terms of affective results, or other results of a more conative than cognitive nature" (Cepollaro, Lepoutre, and Simpson 2022, 7).

If, as Cepollaro, Lepoutre, and Simpson (2022) note, counterspeech does prove to be effective, we are left with the Deontic question: who should be involved in counterspeech, and is there a universal duty to engage in it?

Some authors (for example, Goldberg 2020 and Howard 2021) argue that everyone, provided that it is safe, has a duty to engage in counterspeech just as we, as citizens, have a duty to help when we witness someone hurting another person. These duties would fall on anyone; however, some authors claim that the cost for individuals may be too high and that the duty to engage in counterspeech should fall on institutions (Brettschneider 2012; Gelber 2012; Lepoutre 2017; Saul 2021). As Cepollaro, Lepoutre, and Simpson (2022) stress, on the one hand, the state's response to harmful speech may be a better option to confront hate speech using its various resources. On the other hand, the state might provide resources to individuals, enabling them to engage in counterspeech. This could be done, for example, through funding groups that would counter harmful speech. But, as Cepollaro, Lepoutre, and Simpson (2022) argue, to address the Deontic question, we first need to show that counterspeech is indeed effective; otherwise, our efforts to engage in it would be futile. I will not engage in a more thorough discussion on the issues with counterspeech presented here and I will instead leave it for a future endeavor. However, I think that counterspeech is not to be diminished or rejected and that it could be a useful tool in dealing with harmful speech while preserving freedom of speech. Bearing this in mind, I will proceed by proposing a two-model approach for dealing with harmful speech. Traces of these responses will fall into the category of counterspeech, but as I have stressed in the introduction, if counterspeech alone proves ineffective, this two-model approach could be an appropriate form of action.

5.3. Responses by individual members of society

The responses by individual members of society can be viewed from three perspectives established earlier: the target's, the listener's, and the speaker's. These responses may come from individuals who are engaged in the discourse where the slur was used, or from groups—specifically, target groups—that may use strategies such as appropriation to combat slur use. The responses based on the target's, the listener's, and the speaker's perspective, i.e., the individual/group responses, have one thing in common, and that is epistemic responsibility. Epistemic responsibility, which "concerns our responsibility qua knowers and learners" (Medina 2013, 121), implies cultivating certain epistemic virtues that will help individuals become more epistemically responsible agents. In addition, Medina argues that "responsible agency requires that one be minimally knowledgeable about one's mind and one's life, about the social world and the particular others with whom one interacts, and about the empirical realities one encounters" (Medina 2013, 127). However, he is cautious and elaborates that "cognitive minimums can be assumed to be the norm only insofar as subjects are unimpeded in their processes of knowledge acquisition" (Medina 2013, 129), i.e., only if the conditions of epistemic justice are met. This point will be important for the second set of responses, which I will elaborate on later in the text. For now, I would also add that Fricker (2007) advocates for critical self-reflection, as it may help us recognize prejudice that might be contaminating our belief system. I will now proceed by examining three perspectives, each playing a role in responding to slurs and hate speech. These responses often overlap, especially in the speaker's and the listener's case.

5.3.1. The target's response

I will start with the possible targets' responses because they are the first on the line when it comes to slurs, making it logical to begin with the target's possible reactions. The response of the target will depend on the effect the target is responding to, i.e., whether the target is responding to the primary or the secondary effect.

As mentioned, when the target is facing the primary effect of the slur, their response may resemble the type described by Lawrence (1993), which has speech-inhibiting elements triggering the fight-or-flight response in targets. Lawrence gave an example of one of his students who, upon being targeted by a slur, was "in a state of semi-shock, nauseous, dizzy, unable to muster the witty, sarcastic, articulate rejoinder he was accustomed to making" (Lawrence 1993). In that sense, targets can be silenced by slurs.

Silencing is a notion that was first put forth by feminist authors discussing pornography, most notably MacKinnon (1993), and further developed by other authors

(Langton 1993; Langton and West 1999; Hornsby 1994; Hornsby and Langton 1998; Maitra 2009; McGowan 2004, 2009, 2014; Mikkola 2011, 2019; Caponetto 2021). Luvell and Barnes (2022), in paraphrasing MacKinnon (1993), explained:

> there are some speech acts that fix the possibility of other speech acts. In other words, they make it possible for some persons to perform some speech acts, and make it impossible for others. This is most evident in formal settings, like a legislature, where the formal rules determine who may speak when, and in what manner. Pornography, the argument continues, does just this. It sets rules of behavior that, in effect, inhibit the speech of women. The result of which is that the speech acts of pornography—performed by those who produce and distribute it—create a climate that undermines women's capacity to perform certain speech acts of their own. The speech of some (pornographers), therefore, curtails the speech of others (women). (Luvell and Barnes 2022)

The same line of argument can be applied to hate speech, namely that there are some speech acts that can silence marginalized groups. Luvell and Barnes (2022) have already raised this issue. Hate speech operates by creating an environment that is poisonous to its targets, potentially leading them to withdraw from deliberation.

In a face-to-face exchange, if the target manages to get through the initial shock of facing a slurring comment directed at them, they may opt for confrontation. This is likely the most extreme case in which the slur has incited violence. This is also probably the most clear-cut case of when we could opt for legal sanctioning, and this is also the legal approach taken in the USA with the fighting words doctrine discussed in Chapter I. In that case, we have a right to infringe on one's liberty because their speech caused direct incitement to violence.

However, as argued in Chapter IV, uses of slurs and, more generally, hate speech, may have more systematic and long-lasting effects. For these effects to occur, the utterance need not be said directly to the target; it can be said about them, in absentia. This, in turn, calls for a different response from the targets, and that is appropriation, or reclamation, as some authors refer to it. Appropriation is a form of counterspeech, and counterspeech "is communication that seeks to counteract potential harm that is brought about by other speech" (Cepollaro, Lepoutre, and Simpson 2022, 3). Counterspeech can take many forms and be performed by various agents, including the very targets of harmful speech.

One of the ways targets may engage in counterspeech, addressing the secondary, long-term effects of slurs, is through appropriation, i.e., when slurs are used by members of the target group in a non-derogatory way (such as when *ni**er* is used among members of the black community, or *bitch* among women). Appropriation of slurs has been a long-debated subject in the philosophy of language with various accounts concerning the nature of appropriation. I will not be dealing with these accounts here. Instead, I will focus on one of

the criticisms that is useful for our discussion in this thesis, specifically the issue of the stereotype embedded in the semantics of the slur. Even though negative identity prejudice may be placed either in semantics or pragmatics, I will use this opportunity to raise some questions for authors who propose placing the stereotype in the pragmatics of a slur. But, before I turn my focus on that, I will briefly touch upon what appropriation is. As Bianchi (2014) noted, slurs can be appropriated in two contexts: among friends or in socio-political contexts where the appropriation is used as an attempt to subvert norms. Jeshion (2020), for example, claims that there are two types of slur reclamation: pride reclamation, where target groups express pride for being members of the in-group, and insular reclamation, where the reclamation uses of slurs function to express camaraderie and fight oppression. In any case, and regardless of which theory we endorse, appropriation (or reclamation) of slurs serves as a kind of weapon target groups have at their disposal to fight back. The issue of appropriation (or reclamation) of slurs is a long-debated topic in the philosophy of language, with many unanswered questions. My goal is not to cover all the possible answers or to provide novel ones, but to offer insight into the debate. The points I will make will surely need to be developed further.

Let me now turn my focus to one line of argument regarding the placement of the stereotype within the semantics of a slur. One of the problems Popa-Wyatt and L. Wyatt identify with semanticist theories is the case of appropriation. The problem they see is the fact that the "stereotype is encoded semantically, and as far as semantics is able to explain offence, then the offence caused should be the same, whoever says it" (Popa-Wyatt and L. Wyatt 2018, 2884) which obviously isn't the case with appropriation. Many theorists have suggested that the solution lies in a change in the meaning of the word (Hom 2008; Richard 2008; Potts 2007). However, another problem raised by Anderson and Lepore (2013) is that, if that were the case then any speaker, even a member of the out-group, would be able to use this appropriated meaning (if the meaning has changed into a positive one), which is not the case since using appropriated slurs is reserved for the in-group. To begin addressing this, it is essential to note that language is fluid, constantly changing and adapting. Some words become archaic, new words enter the language, and existing words can acquire new meanings. Slang words are excellent examples of this. For example, the word *tea* has become a slang word meaning to gossip, as in the sentence, "His friends were spilling the tea on what it's like to work for the new boss," meaning they were gossiping about the new boss. In the same fashion, slurs may also gain new meanings, which is something that semantic theorists have argued for. But, Popa-Wyatt and L. Wyatt see two problems with this explanation. First, they echo Anderson and Lepore's sentiments: "why is it the case that only in-group members can access one of these meanings" (Popa-Wyatt and L. Wyatt 2018, 2884)? The reply I wish to offer has two strands: 1) the layered approach to slurs, and 2) the in-group reference. Starting with 1), as already noted in Chapter IV, we should consider slurs as having layers. In that case, when a slur is appropriated, the layer with the "bad" material is peeled off,

namely the negative identity prejudice one, while the negative evaluative one is changed. When a slur is used in an appropriated sense, we essentially peel off the layer that is the negative part of the slur and replace it with a positive meaning (similar account has been made by Zeman 2021). This can be demonstrated with the following:

a) X is a *ni**er*[1].

b) Yeah, damn right he is a *ni**er*[2]!

In a), the *ni**er*[1] signifies a literal meaning where the author conveys that X is:

layer 1) a black person,

layer 2) dirty, lazy, unintelligent, and so on because of being black,

layer 3) negative identity prejudice that black people are lazy, unintelligent, violent, prone to anger, and so on,

layer 4) historical link to the time members of the target group were labeled with the slur,

layer 5) a feeling of contempt and disgust towards members of the target group,

layer 6) epistemic (non)culpability.

In b), the slur is used in an appropriated sense, and the speaker conveys that X is:

layer 1) a black person,

layer 2) strong, brave, unapologetic, and so on because of being black,

layer 3) positive identity prejudice that people of color are proud of who they are and of their heritage, and that they are strong, independent individuals,

layer 4) historical link to the time members of the target group were labeled with the slur,

layer 5) a feeling of pride towards members of the target group,

layer 6) epistemic (non)culpability.

Notice that, in the appropriated case, certain layers have changed their meaning. The negative identity prejudice present in the first case is now replaced with positive identity prejudice, which in turn generates a positive evaluative layer. Negative identity prejudice associated with being black is no longer in use in the appropriated sense because the word is changing. Once the appropriation process is complete, the word will have changed its layers to create an entirely new meaning, such is the case with *queer*, for example.

Continuing with the second line of argument, which is the in-group reference argument, the appropriated use of *bitch* would have the meaning of "we are proud to be women who know what they want". I would argue that all appropriated uses have this self-in-group reference in the meaning (so, for example, *ni**er* in its appropriated sense might convey something like "I'm black and I'm proud!" or the plural variation "We're black and we're proud!"). In that case, it doesn't make much sense that an out-group member uses it since they aren't a woman/black, i.e., the neutral counterpart of the word doesn't apply to them. The second question Popa-Wyatt and L. Wyatt ask is "what happens during the process of appropriation" because, they wonder, "if derogation resides in semantic meaning, and not pragmatic effects, then how can Hom explain that 'queer' can be used both derogatorily and non-derogatorily during the reclamation process" (Popa-Wyatt and L. Wyatt 2018, 2884)? First, during the process of appropriation, the word only peels off layers, as demonstrated in an earlier example. Second, after appropriation, the word acquires a new meaning because all of the layers have been replaced with new ones. However, as seen in dictionaries, words often have multiple meanings, so sometimes it will be the case that more than one meaning of the word is used. Despite the issues with the reclamation process I have just outlined, it remains one of the possible ways for the targets to take control of the narrative.

5.3.2. The speaker's and the listener's possible responses

It is probably uncommon to review the possible responses the speaker might have to what has been uttered, and to the best of my knowledge, not much has been said in the literature about it. Nonetheless, I believe it is worth examining what kind of responses the speakers might have, even if it is to their own speech. Some of these responses will align with the responses of the listener. The responses I have in mind are virtues, i.e., what kind of virtues one must cultivate in order to avoid being susceptible to using and endorsing hate speech. But, first, I would like to divide these responses into two groups: direct and indirect responses. Direct responses would be responses that take place at the moment of utterance and in a face-to-face setting. Indirect responses would be the ones taking place independently of the utterance, not in a face-to-face setting, but remotely, so to speak.

Listeners have an array of possible direct responses at their disposal. When encountered with hate speech, one of the responses of the listeners could be their agreement with what was uttered, thereby aligning themselves with the speaker. They might agree with what was said for a number of reasons: maybe they are sexist or homophobic themselves, maybe they feel pressured by the speaker's position of power over them, and so on. For whatever reason, their response is agreement which they can verbally express by giving support to the speaker. Another possible response is silence. Here, it is important to differentiate between silence as agreement and silence without agreement. Additionally,

listeners might disagree with what was said and verbally express their disagreement, engaging in counterspeech whereby they can morally condone the utterance by verbally expressing themselves or by sanctioning the speaker by removing themselves from their company or severing any future social contact with the speaker.

First of all, I would like to reiterate my point made in Chapter IV about the (non)culpability of the speaker. Once the speaker has uttered a derogatory word, we can imagine they will, at some point, be presented with counterevidence. Once that happens, the speaker has two choices: they can retract and admit their mistake, and adjust their belief system accordingly (presupposing, of course, that they were unfamiliar with what they were saying and conveying), or they can resist the counterevidence presented. In the former case, we might consider them to be epistemically non-culpable; they have made an honest mistake. In the latter case, the speaker is epistemically culpable because they refused to change their belief system.

In Chapter IV, I have mentioned that in some cases certain speakers are susceptible to greater scrutiny than others, namely because of the role they have in a community. An example I gave was that of politicians who are elected representatives of the people and due to their role, they succumb to greater scrutiny when it comes to the way they express themselves in public because, as I argued, their words can have a greater perlocutionary effect (as seen in the Trump example). That's why epistemic responsibility is an important aspect of our social lives. Since we are part of a community, we have certain obligations towards our fellow citizens, and I think one of those obligations is being epistemically responsible agents. In order to work towards being epistemically responsible agents, we need to cultivate some epistemic virtues as well. These virtues would potentially help both speakers and listeners to resist hate speech.

Fricker (2007), when discussing testimonial sensibility, which she defines as "an idea of a spontaneous critical sensitivity that is permanently in training and continuously adapting according to individual and collective experience" (Fricker 2007, 84), emphasizes that one needs to have a particular virtue in order to account for the prejudices one may have or encounter. These virtues Fricker discussed may be one of the ways individuals can use to counteract the prejudice they may have or encounter. Even though Fricker's focus is on testimonial exchanges, we can extrapolate her conclusions to account for how prejudice may affect other aspects of our social lives as well. The virtue she has in mind is reflexive critical awareness of the possible prejudice one may hold, and I argue that this virtue she mentions may be applied to counteract prejudice on both the speaker's and the listener's parts. Fricker explains what it means to correct for prejudice by using reflexive critical awareness as follows:

> When the hearer suspects prejudice in her credibility judgment – whether through sensing cognitive dissonance between her perception,

> beliefs, and emotional responses, or whether through self-conscious reflection – she should shift intellectual gear out of spontaneous, unreflective mode and into active critical reflection in order to identify how far the suspected prejudice has influenced her judgment. (Fricker 2007, 91)

This way, explains Fricker, any negative impact of prejudice should be neutralized. In some cases, we may need to suspend judgment altogether or seek further evidence. Fricker's explanation of this reflexive critical awareness can be applied to correcting for prejudice in general, beyond just testimonial exchanges. By being able to have an insight into one's own beliefs and the prejudice that may influence those beliefs, one could be aware of the negative impact slurs carry with them. It would also mean one would be able to better recognize the identity prejudice that is evoked when a slur is uttered. This could be applied to both the speaker and the hearer. Reflexive critical awareness could be crucial not only for developing testimonial sensibility but also for developing epistemic responsibility more generally.

Similarly to Fricker, Medina (2013) argues that:

> ...epistemic responsibility involves, crucially (perhaps even constitutively), obligations to know oneself and to know others with whom one's life and identity are bound up. In order to acquire and transmit knowledge responsibly, one needs to be critically aware of one's identity and that of others; one must have at least a minimal amount of self-knowledge and social knowledge of others. The exact kind and amount of self-knowledge and knowledge of others required for responsible agency can only be contextually determined, taking into account who one is, the kinds of epistemic actions and transactions in which one engages, and the socio-historical contexts in which one lives. (Medina 2013, 54-55)

Following the discussion about the epistemic vices of the privileged, Medina also analyzes the virtues that are characteristic of the oppressed subjects, namely humility, curiosity/diligence, and open-mindedness. But Medina also acknowledges that these virtues are not only exclusive to the oppressed, nor are they automatic; subjects do not simply have them by being members of oppressed groups. Building on this, I hold that these virtues could be helpful if developed by the speakers and the listeners. The first virtue Medina describes is epistemic humility, which is being aware of one's cognitive limitations. The idea is that such a virtue may facilitate cognitive improvement and some learning processes. When it comes to the speaker and the listener, adopting a self-questioning attitude could be beneficial when one is faced with counterevidence. First, if a speaker is epistemically humble, this would mean that they are more prone to accepting the counterevidence they are presented with. On the other hand, listeners could utilize this virtue in order to question their own prejudices and

become aware of how these prejudices might hinder their judgment of the targets. The second virtue Medina mentions is intellectual curiosity/diligence, which motivates a subject to fill their cognitive gaps, once they come to know them. This is a virtue that can, once more, be beneficial to both speakers and listeners. If the speakers and listeners possess this virtue, that means they are eager to learn and acquire knowledge which would lead to filling cognitive gaps and, hopefully, accounting for prejudice they may hold. The last virtue Medina writes about is open-mindedness. Medina notes how open-mindedness is not characteristic of the privileged (in our case the privileged would be the speakers, i.e., users of slurs) because they rarely show the willingness or ability to see and acknowledge other perspectives. Bearing this in mind, it would be more than beneficial for speakers to be able to acquire this virtue, which would make them more open to other viewpoints, i.e., the viewpoints of the underprivileged minorities they target. This way, they could become aware of the potential harm they might inflict on them. All of the mentioned virtues would work towards becoming a more epistemically responsible agent. Listeners would be more attuned to their own potential prejudices and the prejudices surrounding them, making them less susceptible to the influence of slurring terms. Speakers, on the other hand, would be more open to accepting counterarguments and possibly more willing to adjust their belief systems. However, there are two issues that complicate things. First, in order to have the virtues mentioned above, individuals must have the opportunity to develop them, which requires living in a society that provides and secures such opportunities. In other words, society needs to provide institutionalized support to enable its citizens to develop these virtues. This is something I will be dealing with in the next section. The second issue concerns the unconscious processes already discussed. As shown in Chapters II and IV, there are some implicit biases[71] that we may harbor even if we are not fully aware of them. That notion could make the reflexive critical awareness difficult, if not impossible, to achieve. To reiterate some of the points made earlier, Moles argues how mental contamination could be detrimental even to our autonomy. As Moles notes, some external forces, namely society which is riddled with stereotypes and prejudice, can lead to responses and unconscious processes that we may not be able to identify. So, for example, a person who is not racist may have some racist responses, despite their belief system not being in accordance with racist ideology. Similar observations are made by Fricker (2007), who claims that stereotypes can influence our judgments and behavior without our conscious awareness. There is empirical research that can corroborate these claims about unconscious processes (Devine 1989; Correll and colleagues 2002). This suggests that it would be an (unattainable?) epistemic endeavor to successfully account for prejudices that linger in our minds. I do not believe that reflexive critical awareness, or other virtues I mentioned, are entirely futile efforts. However, given the unconscious processes that may riddle our minds, I think these virtues do require external support. As already

[71] See Saul 2017.

foreshadowed, this external support would need to come from institutions that would work towards creating an environment in which each person is provided with greater opportunities to develop into an epistemically responsible agent.

5.4. Institutional responses

As previously explained, relying solely on virtues that should somehow be developed in targets, speakers, and listeners is not sufficient for various reasons.

First, due to implicit bias and stealth mode of stereotypes (Saul 2013, 2017; Fricker 2007; Prijić-Samaržija 2020), the success of acquiring those virtues, and consequently the successful elimination of stereotypes in our judgment, is questionable. That being said, I do not think that pursuing these virtues is a futile endeavor. But, in order to acquire them, and later on make use of them, one needs institutionalized support. People tend to be cognitively biased, prone to forming epistemic bubbles that echo their own values, and those who hold negative identity prejudice tend to be resistant to counterevidence. In order to avoid that, institutionalized support must be established as a foundation for developing one's virtuous character. Even though we may feel, perhaps correctly, that forming such virtues that would completely account for our prejudice is an impossible task, it is still worthwhile to get at least a little bit closer to the goal of having virtues that would help us become epistemically more responsible agents. In other words, institutionalized support is a precondition to developing said virtues. On the other hand, if we feel that this task is unattainable, then having institutionalized support becomes necessary in order to account for prejudice we might not be able to recognize or overcome on our own.

Second, placing the burden of fighting hate speech on individuals is a demanding task, especially when the targets are left to defend themselves. Therefore, I argue that the state has a responsibility to respond to the harms inflicted on its most vulnerable members by providing certain institutionalized protections.

Samaržija and Cerovac (2021) have provided considerable insight into what kind of institutionalized measures would have to be taken in order to ameliorate the distributive form of epistemic injustice:

> To remedy epistemic injustice, we must transform the social environment where these transactions take place. In Estlund's terms, we are not dealing with aspirational justice, concerned with defining what is just in ideal conditions of thorough agential virtue, but with concessive justice, which seeks to divulge workable solutions to social problems (Estlund 2020). While we explicitly locate the liability in agents and structures harbouring negative identity prejudices, we acknowledge the empirical evidence suggesting that pleas for individual virtue will likely fail to be realized in practice. Unlike aspirational proposals, such as Fricker's appeal for agential epistemic virtue, concessive justice recognizes sober obstacles to achieving a fully just society, such as citizens' biases and other contingent psychological factors. (Samaržija and Cerovac 2021, 2)

Thus, they propose a set of four institutionalized measures that would help remedy epistemic injustice that plagues our social world. Samaržija and Cerovac are motivated by the fact that the proposed virtue model has not been entirely successful in accounting for a) implicit bias where it was shown that self-reflecting can backfire (Saul 2017), b) distributive injustice in education, i.e., the fact that marginalized groups tend to lack the training to display expertise due to them being discouraged from pursuing higher education, and c) epistemic objectification (Haslanger 2017) where one may form a false belief that negative prejudices are rooted in nature (Samaržija and Cerovac 2021). The best way forward is, according to Samaržija and Cerovac (2021), to change the social environment in which prejudicial thinking arises. I agree with their conclusion and hold that providing certain institutionalized measures that could have the effect of ameliorating and correcting our biases from the start is a fruitful effort. Having this kind of support will in turn also foster virtues that could help us become epistemically more responsible.

In order to better our epistemic environment, Samaržija and Cerovac propose four measures, although these are not necessarily exhaustive. Two of the measures are more immediate, while the other two will require a more systematic approach.

First, to develop our capabilities as knowers, we need to have fair access to education. Unfortunately, this is not the case for everyone, as some lack such access. This distributive form of epistemic injustice "is in place when a necessary epistemic resource – quality education – is so unfairly distributed that marginalized groups do not fulfill the minimum of their epistemic potential" (Samaržija and Cerovac 2021, 10). This unfair access creates a vicious circle: when quality education is unattainable for marginalized members due to factors like gender, ethnicity, or class, their contribution to public discourse becomes epistemically inferior. When this becomes statistically regular, it may reinforce negative identity prejudice. According to Samaržija and Cerovac (2021), what is needed is access to quality education, along with affordable housing and other remedies that could even the ground for marginalized groups.

Second, voicing our perspectives to others is one way to bolster epistemic justice which requires "open political, cultural, and academic practices where disparate social groups can interact as equals" (Samaržija and Cerovac 2021, 10). Instead of voicing their issues solely within closed communities, open political, social, and media platforms would enable these groups to share their problems with the rest of society, thus enriching it with new perspectives and vocabulary, potentially bolstering understanding of the issues they face. These platforms could also provide counterevidence to ongoing prejudice and help neutralize hermeneutical injustice.

Third, following the perspective of concessive justice, another solution to reduce prejudicial thinking is to provide means for marginalized individuals to be able to reach public offices. As Samaržija and Cerovac (2021) explain their point:

> The aim of fair access to public posts is, first, to offset the identity prejudice that excludes marginalized knowers from responsible work despite the quality of their contributions. Second, it refutes negative prejudices about their epistemic potential. In the past, some authors have been doubtful of initiatives that might impose additional regulations upon science, as it is already burdened by incentive structures intending to render it more productive. Yet, we hold that institutionalized aid for marginalized groups balances the fact scientists are fallible agents who also fall prey to socially shared prejudice. (Samaržija and Cerovac 2021, 11-12)

Samaržija and Cerovac (2021) list three potential benefits of this approach: First, once in the right positions, marginalized groups could improve their group's unfavorable image, and by being successful in those positions, they could provide counterevidence to existing identity prejudice. Second, this could neutralize the notion that inequalities are rooted in nature, and instead, marginalized groups could use their positions to foster a more positive image of their groups. Third, by filling important decision-making roles, such as juries and hiring committees, marginalized individuals would provide dominant groups with additional epistemic resources, potentially reducing hermeneutical ignorance.

Finally, Samaržija and Cerovac, (2021) call for anonymous reviews and standardized performance assessments that could alleviate individuals' biases towards marginalized groups. They argue that these measures could be particularly useful in certain competitive business or academic settings, where biases might otherwise result in judgments based on group membership rather than on individual competence.

Samaržija and Cerovac (2021) have offered compelling and fruitful options that institutions have at their disposal to alleviate epistemic injustices stemming from negative identity prejudice. I agree with their approach and think that insisting on developing certain virtues without systematic support from institutions is a fruitless effort. Even though some may argue how hoping that the virtue approach will help us correct for prejudice we may hold is futile, I think that providing the intended institutionalized support may have the desired effect of forming virtues that would alleviate (at least to some degree) some prejudice we have or may encounter. Relying on virtues alone may be insufficient due to the challenges we mentioned earlier, such as implicit bias. However, with proper institutional support to help us develop these virtues, we would be taking a step in the right direction.

The kind of institutional responses we have just discussed could be viewed as preventive measures and prerequisites for advancing epistemic justice and epistemic responsibility. This means that the set of these measures could help us prevent (to a degree) certain prejudices from circulating further in society, i.e., these measures could lessen the influence of these prejudices because we would be faced with counterevidence and we would be better equipped for handling them once we do encounter them. This can be achieved by

providing access to quality education for all, with an emphasis on teaching empathy towards our fellow citizens (as well as people in general) and providing enough counterevidence early on in order to avoid creating and/or perpetuating prejudice against certain groups. But, as Samaržija and Cerovac (2021) have correctly noticed, providing education alone would not be enough if one's basic needs, such as housing, food, and the like, are not met. So, in order to ensure that individuals can reach their full potential as epistemic agents in academic setting and beyond, institutions must first make sure that these fundamental needs are met. Only then will the education that teaches empathy and combats prejudice be truly effective.

However, combating the prejudices that produce the harms discussed in Chapter IV and create derogatory-labeling injustice requires immediate action. Therefore, alongside preventive measures, there is also a need for correction. By this, I mean that in certain cases, the government may need to respond with legal sanctions. When addressing hate speech, there are instances that necessitate legal sanctioning. Many countries already have laws in place for this purpose, and we have discussed and reviewed some of them in Chapter I. Freedom of speech is not absolute and there are cases when it can be infringed. These cases, however, vary from country to country. As foreshadowed earlier in Chapter IV, derogatory-labeling injustice could be a reason for legal action, particularly in cases where it overlaps with hate speech. This is not to suggest that the state should respond with legal bans *only* in cases where there is an overlap between derogatory-labeling injustice and hate speech. However, given the lack of a unified definition of hate speech and the difficulties courts often face in determining which language to legally sanction, incorporating derogatory-labeling injustice into the picture provides a clearer way of distinguishing when it would be acceptable to resort to legal sanctions. Of course, since derogatory-labeling injustice specifically pertains to cases involving slurs, there may be other forms of hate speech that also warrant legal sanctioning, depending on our accepted definition of hate speech. It is not my intention here to delve into the discussion of which cases would warrant such action; rather, I aim to highlight that institutions have that option available and that it should be used in cases where hate speech overlaps with derogatory-labeling injustice. In addition, even in cases where legal sanctions are applied, they do not preclude the use of counterspeech.

In cases where there are no strong enough reasons for legal bans—such as when derogatory-labeling injustice occurs but does not overlap with hate speech—an appropriate institutional response would be to utilize counterspeech. In these cases, while legal restrictions may not be warranted, it remains important to address the derogatory speech causing derogatory-labeling injustice. If that is the case, institutions may employ some of the alternative methods to address the issue by utilizing counterspeech.

CONCLUSION

The aim of the thesis was to provide a deeper understanding of what slurs do when uttered, i.e., to explore the nature of the harm slurs cause and to investigate the underlying mechanisms which enable such harm to occur. To accomplish this, I delved into the interplay between hate speech, slurs, prejudice and the potential harm these phenomena may cause, thus combining and providing an outlook to these issues from various disciplines, namely philosophy of language, political philosophy and epistemology.

One of the main concerns was to further explore the issue of slurs that do not legally qualify as hate speech. Namely, there are some slurs that do not fit into the category of hate speech, but they still cause significant harm to the target. The prime examples of these slurs are gendered slurs for women, such as *whore*, *slut* etc. I claim that the common denominator of what Jeshion (2013a) described as weapon-uses of slurs is the underlying negative identity prejudice identified by Fricker (2007). These tracker prejudices follow us through every social aspect of our lives. Thus, when uttering a slur in its literal sense to degrade, these negative identity prejudices are evoked. This gives rise to a novel kind of injustice: derogatory-labeling injustice which I defined as occurring in a discourse setting when a speaker uses slurs and thus evokes identity prejudices about the target causing harm to one or more of their important interests. For derogatory-labeling injustice to occur, the target must be a member of a historically marginalized group and the speaker has to hold a certain amount of social power. This novel concept gives us an explanatory advantage: we can now define what kind of harm slurs that are not considered hate speech do, and we can better implement strategies to counter such speech.

To research these issues, I started in the first Chapter by providing an overview of the debate between hate speech and freedom of speech. First, I established that there is no universal definition of hate speech, so legal treatment of hate speech varies through countries, but, nevertheless there are some characteristics that can be extrapolated: that it is public speech and that it targets individuals and groups with ascribed characteristics. In this Chapter I presented arguments from prominent authors who argue for protection of hate speech and from authors who argue for restricting hate speech. I also juxtaposed the legal treatment in the US and Croatia to show how different understanding of the definition of hate speech transfers to the different treatment of hate speech in the legal domain.

In the second Chapter, I proceeded to build a needed background on how stereotypes and prejudice may affect us. I presented various empirical research on the effect of stereotypes and prejudice, as well as derogatory language, in order to trace back the harm caused by derogatory language, i.e., slurs to prejudices.

In the third Chapter, I addressed the most used vehicle of hate speech – slurs. First, I untangled the concerns raised by Nunberg (2018) and Ashwell (2016) regarding gendered

slurs for women. This was an important step because it seems that gendered slurs for women present prime examples of slurs that wouldn't normally be considered hate speech but that still cause harm usually accredited to hate speech. I endorsed the view presented by Legaspe (2018) and strengthened it by proposing that slurs have an underlying negative identity prejudice that is evoked every time a slur is used in its literal sense to degrade, and that these prejudices apply to all members of the group not just the individual. This, in turn, is also the main augmentation I made to the existing theories of slurs. Namely, by using Miščević's (2016) account of layers of slurs, I introduced negative identity prejudice as being part of the slur which provides us with an explanatory advantage of a slur's content: the normative evaluative judgment is grounded in negative identity prejudice.

After setting up a needed background from philosophy of language, in the fourth Chapter I turned to the pivotal aspect of the thesis and introduced the harm caused by slurs and the novel kind of injustice: derogatory-labeling injustice. Derogatory-labeling injustice happens in a discourse setting where the speaker holds a certain amount of social power and uses slurs in their literal sense to target historically marginalized groups thus harming one or more of their important interests. These interests are: a) being successful in one's academic life which is a foundation for economic stability later in life and career opportunities (which can, as we have seen, also be influenced by stereotype threat); b) being able to form one's identity free of the kind of influences due to which we can form a negative image of ourselves or our role in a community; c) being able to hold a social standing in a community that will enable us to have equality of opportunity; d) being able to participate in deliberation equally; e) being able to acquire primary goods; f) being able to pursue interests that will enable us to communicate our thoughts to others so that others may come to know us as an individual that we are. All of the mentioned interests may be obstructed by the systematic use of negative identity prejudice directed at targets via slurs. Finally, we are better able to understand the demarcation between hate speech, slurs and derogatory-labeling injustice where some slurs which don't legally classify as hate speech cause derogatory-labeling injustice and thus warrant our attention.

In the fifth Chapter, I deal with possible answers to such phenomena. I conclude that we have good reasons to legally restrict speech that classifies as hate speech and that causes derogatory-labeling injustice. Moreover, we also have reasons to sanction and address speech that isn't hate speech but that causes derogatory-labeling injustice. These sanctions would not need to be legal ones, they can come in the form of counterspeech or a moral reprimand. Additionally, identifying derogatory-labeling injustice shows us that there is a pressing need to foster a more inclusive and equitable society. I propose this could be done by combining two fronts. On the one hand, we could foster epistemic responsibility in individuals. However, this would need to be supported institutionally where the state should provide

resources that can help individuals develop certain epistemic virtues that can, in the long run, help them recognize stereotypes and prejudice at play in society.

BIBLIOGRAPHY

1. Aboud, Frances E. *Children and Prejudice.* Oxford: Basil Blackwell, 1988.
2. Alaburić, Vesna. "Ograničavanje 'Govora Mržnje' u Demokratskome Društvu - Teorijski, Zakonodavni i Praktički Aspekti." *Hrvatska pravna revija* (2003).
3. Alcoff, Linda Martin. "On Judging Epistemic Credibility: Is Social Identity Relevant?" In *Women of Color and Philosophy,* edited by Naomi Zack, 133–152. Oxford: Blackwell, 2000.
4. Amodio, David M., Eddie Harmon-Jones, Patricia G. Devine, John J. Curtin, Sigan L. Hartley, and Alison E. Covert. "Neural Signals for the Detection of Unintentional Race Bias." *Psychological Science* 15, no. 2 (2004): 88–93.
5. Anderson, Luvell, and Ernie Lepore. "Slurring Words." *Noûs* 47, no. 1 (2011): 25-48.
6. Anderson, Luvell, and Michael Barnes. "Hate Speech." In *The Stanford Encyclopedia of Philosophy* (Fall 2023 Edition), edited by Edward N. Zalta and Uri Nodelman. https://plato.stanford.edu/archives/fall2023/entries/hate-speech/.
7. Anti-Discrimination Act. 2012. *Official Gazette* 112/2012, no. 2430. Croatian Parliament. Published on October 11, 2012.
8. Anzures, Gizelle, Paul C. Quinn, Olivier Pascalis, Alan M. Slater, and Kang Lee. "Development of Own-Race Biases." *Visual Cognition* 21, no. 9–10 (2013): 1165–1182. doi:10.1080/13506285.2013.821428.
9. Aronson, Elliot, Timothy D. Wilson, Robin M. Akert, and Samuel R. Sommers. *Social Psychology*. 9th ed. London: Pearson Education, 2015.
10. Aronson, Joshua, and Michael Inzlicht. "The Ups and Downs of Attributional Ambiguity: Stereotype Vulnerability and the Academic Self-Knowledge of African American College Students." *Psychological Science* 15, no. 12 (2004): 829-836.
11. Arpaly, Nomy. *Unprincipled Virtue: An Inquiry into Moral Agency.* Oxford: Oxford University Press, 2003.
12. Ashwell, Lauren. "Gendered Slurs." *Social Theory and Practice* 42, no. 2 (2016): 228-239.
13. Austin, John L. *How to Do Things with Words*. Oxford University Press, 1962.

14. Baccarini, Elvio. “Liberalni Nacionalizam: Argument Samopoštovanja.” *Filozofska istraživanja* 30, no. 1-2 (2010): 295-310. https://hrcak.srce.hr/62980.
15. Bach, Kent. “Loaded Words: On the Semantics and Pragmatics of Slurs.” In *Bad Words: Philosophical Perspectives on Slurs,* edited by David Sosa, 60-76. Oxford, United Kingdom: Oxford University Press, 2018.
16. Baird, Abigail A., Staci A. Gruber, Deborah A. Fein, Luis C. Maas, Ronald J. Steingard, Perry F. Renshaw, Bruce M. Cohen, and Deborah A. Yurgelun-Todd. “Functional Magnetic Resonance Imaging of Facial Affect Recognition in Children and Adolescents.” *Journal of the American Academy of Child & Adolescent Psychiatry* 35 (1999): 195.
17. Baker, C. Edwin. *Human Liberty and Freedom of Speech.* New York: Oxford University Press, 1989.
18. Beauvoir, Simone de. *The Second Sex.* London: Johnatan Cape, 1949.
19. Ben-Zeev, Talia, Steven Fein, and Michael Inzlicht. “Arousal and Stereotype Threat.” *Journal of Experimental Social Psychology* 41 (2005): 174–181.
20. Bianchi, Claudia. “Slurs and Appropriation: An Echoic Account.” *Journal of Pragmatics* 66 (2014): 35–44.
21. Bianchi, Mauro, Andrea Carnaghi, Valentina Piccoli, Marta Stragà, and Davide Zotti. “On the Descriptive and Expressive Function of Derogatory Group Labels: An Experimental Test.” *Journal of Language and Social Psychology* 38, no. 5–6 (2019): 756–772. https://doi.org/10.1177/0261927X19867739.
22. Bilewicz, Michał, and Wiktor Soral. “Hate Speech Epidemic: The Dynamic Effects of Derogatory Language on Intergroup Relations and Political Radicalization.” *Political Psychology* 41, no. 1 (2020): 3–33.
23. Bhagwat, Ashutosh, and James Weinstein. “Freedom of Expression and Democracy.” In *The Oxford Handbook of Freedom of Speech,* edited by Adrienne Stone and Frederick Schauer, 82-105. Oxford: Oxford University Press, 2021.
24. Blascovich, Jim, Steven Spencer, Diane Quinn, and Claude Steele. “African Americans and High Blood Pressure: The Role of Stereotype Threat.” *Psychological Science* 12, no. 3 (2001): 225-229.

25. Blum, Lawrence. “The Too Minimal Political, Moral, and Civic Dimension of Claude Steele's ‘Stereotype Threat’ Paradigm.” In *Implicit Bias and Philosophy, Vol. 2: Moral Responsibility, Structural Injustice, and Ethics,* edited by Michael Brownstein and Jennifer Saul, 147–172. New York: Oxford University Press, 2016.
26. Bolinger, Renée Jorgensen. “The Pragmatics of Slurs.” *Noûs* 51, no. 3 (2017): 439-462.
27. Bonotti, Matteo, and Jonathan Seglow. “Freedom of Expression.” *Philosophy Compass* 16, no. 7 (2021): e12759.
28. Bosson, Jennifer, Ethan Haymovitz, and Elizabeth Pinel. “When Saying and Doing Diverge: The Effects of Stereotype Threat of Self-Reported Versus Non-Verbal Anxiety.” *Journal of Experimental Social Psychology* 40, no. 2 (2004): 247-255.
29. Brandeis, Louis D. *Whitney v. California.* 274 U.S. 357, 1927.
30. Brandon, C. Welsh, Anthony A. Braga, and Gerben J. N. Bruinsma. “Reimagining Broken Windows: From Theory to Policy.” *Journal of Research in Crime and Delinquency* 52, no. 4 (2015): 447–463. https://doi.org/10.1177/0022427815581399.
31. Brettschneider, Corey. *When the State Speaks, What Should It Say?* Princeton, NJ: Princeton University Press, 2012.
32. Brink, David O. “Millian Principles, Freedom of Expression, and Hate Speech.” *Legal Theory* 7, no. 1 (2001): 119–157.
33. Brown, Alexander. *Hate Speech Law: A Philosophical Examination.* Routledge, 2015.
34. Brown, Alexander. “Hate Speech Laws, Legitimacy, and Precaution: A Reply to James Weinstein.” *Constitutional Commentary* 32 (2017): 599–617.
35. Brown, Rupert. *Prejudice: Its Social Psychology.* Wiley-Blackwell, 2010.
36. Cameron, Jessica A., Jeannette M. Alvarez, Diane N. Ruble, and Andrew J. Fuligni. “Children’s Lay Theories about Ingroups and Outgroups: Reconceptualizing Research on Prejudice.” *Personality and Social Psychology Review* 5 (2001): 118–128.
37. Camp, Elisabeth. “Slurring Perspectives.” *Analytic Philosophy* 54, no. 3 (2013): 330-349.

38. Caponetto, Laura. “A Comprehensive Definition of Illocutionary Silencing.” *Topoi* 40 (2021): 191–202.
39. Carnaghi, Andrea, Anne Maass, and Fabio Fasoli. “Enhancing Masculinity by Slandering Homosexuals: The Role of Homophobic Epithets in Heterosexual Gender Identity.” *Personality and Social Psychology Bulletin* 37, no. 12 (2011): 1655–1665. https://doi.org/10.1177/0146167211424167.
40. Castelli, Luigi, Cristina De Dea, and Drew Nesdale. “Learning Social Attitudes: Children’s Sensitivity to the Nonverbal Behaviors of Adult Models During Interracial Interactions.” *Personality and Social Psychology Bulletin* 34 (2008): 1504–1513.
41. Castelli, Luigi, Cristina Zogmaister, and Silvia Tomelleri. “The Transmission of Racial Attitudes within the Family.” *Developmental Psychology* 45 (2009): 586–591.
42. Cepollaro, Bianca, Maxime Lepoutre, and Robert Mark Simpson. “Counterspeech.” *Philosophy Compass* 18, no. 1 (2022): e12890.
43. Cervone, Carmen, Martha Augoustinos, and Anne Maass. “The Language of Derogation and Hate: Functions, Consequences, and Reappropriation.” *Journal of Language and Social Psychology* 40, no. 1 (2021): 80–101.
44. Cikara, Mina, and Jay J. Van Bavel. “The Neuroscience of Intergroup Relations: An Integrative Review.” *Perspectives on Psychological Science* 9 (2014): 245–274.
45. Clark, Kenneth B., and Mamie P. Clark. “Racial Identification and Preference in Negro Children.” In *Basic Studies in Social Psychology,* edited by H. Proshansky and B. Seidenberg, 308–317. New York: Holt Rinehart and Winston, 1947.
46. Cohen, Joshua. “Freedom of Expression.” *Philosophy & Public Affairs* 22, no. 3 (1993): 207–263.
47. Correll, Joshua, Bernadette Park, Charles M. Judd, and Bernd Wittenbrink. “The Police Officer’s Dilemma: Using Ethnicity to Disambiguate Potentially Threatening Individuals.” *Journal of Personality and Social Psychology* 83 (2002): 1314–1329.
48. Council of Europe. *Recommendation No. R (97) 20.* 1997.
49. Crano, William D., and Phyllis M. Mellon. “Causal Influence of Teachers’ Expectations on Children’s Academic Performance: A Cross-Lagged Panel Analysis.” *Journal of Educational Psychology* 70, no. 1 (1978): 39–49.

50. Criminal Code. 2011. *Narodne novine* (Official Gazette) 125/2011, no. 2498. Croatian Parliament. Published on November 7, 2011.
51. Davey, Alfred. *Learning to Be Prejudiced: Growing Up in Multi-ethnic Britain.* London: Edward Arnold, 1983.
52. Davies, Paul, Steven Spencer, Diane Quinn, and Rebecca Gerhardstein. "Consuming Images: How Television Commercials that Elicit Stereotype Threat Can Restrain Women Academically and Professionally." *Personality and Social Psychology Bulletin* 28, no. 12 (2002): 1615-1628.
53. Davies, Paul, Steven Spencer, and Claude Steele. "Clearing the Air: Identity Safety Moderates the Effects of Stereotype Threat on Women's Leadership Aspirations." *Journal of Personality and Social Psychology* 88, no. 2 (2005): 276-287.
54. Davis, Christopher, and Elin McCready. "The Instability of Slurs." *Grazer Philosophische Studien* 97, no. 1 (2020): 63-85.
55. Devine, Patricia G. "Stereotypes and Prejudice: Their Automatic and Controlled Components." *Journal of Personality and Social Psychology* 56, no. 1 (1989): 5–18.
56. Devine, Patricia G., and Steven J. Sherman. "Intuitive Versus Rational Judgement and the Role of Stereotyping in the Human Condition: Kirk or Spock?" *Psychological Inquiry* 3 (1992): 153–159.
57. Devine, Patricia G., E. Ashby Plant, David M. Amodio, Eddie Harmon-Jones, and Stephanie L. Vance. "The Regulation of Explicit and Implicit Race Bias: The Role of Motivations to Respond Without Prejudice." *Journal of Personality and Social Psychology* 82, no. 5 (2002): 835–848.
58. Douglas, Karen M. "Psychology, Discrimination and Hate Groups Online." In *The Oxford Handbook of Internet Psychology,* edited by A. Johnson, K. McKenna, T. Postmes, and U. Reips, 155–164. Oxford: Oxford University Press, 2007.
59. Dworkin, Ronald, ed. *Freedom's Law: The Moral Reading of the American Constitution.* Oxford: Oxford University Press, 1996.
60. Dworkin, Ronald. "Foreword." In *Extreme Speech and Democracy,* edited by Ivan Hare and James Weinstein, v–ix. Oxford: Oxford University Press, 2009.

61. Eccles, Jacquelynne S., Janis E. Jacobs, and Rena D. Harold. “Gender Role Stereotypes, Expectancy Effects, and Parents’ Socialization of Gender Differences.” *Journal of Social Issues* 46 (1990): 183–201.
62. Estlund, David. *Utopophobia: On the Limits (If Any) of Political Philosophy.* Princeton: Princeton University Press, 2020.
63. European Convention on Human Rights. *Article 10.* 1950.
64. Fasoli, Fabio, Anne Maass, and Andrea Carnaghi. “Labelling and Discrimination: Do Homophobic Epithets Undermine Fair Distribution of Resources?” *British Journal of Social Psychology* 54, no. 2 (2015): 383–393. https://doi.org/10.1111/bjso.12090.
65. Fasoli, Fabio, Maria Paola Paladino, Andrea Carnaghi, Jolanda Jetten, Brock Bastian, and Paul G. Bain. “Not ‘Just Words’: Exposure to Homophobic Epithets Leads to Dehumanizing and Physical Distancing from Gay Men.” *European Journal of Social Psychology* 46, no. 2 (2016): 237-248. https://doi.org/10.1002/ejsp.2148.
66. Feinberg, Joel. *Harm to Others. Volume 1, The Moral Limits of the Criminal Law.* Oxford: Oxford University Press, 1984.
67. Feinberg, Joel. *Offense to Others. Volume 2, The Moral Limits of the Criminal Law.* Oxford: Oxford University Press, 1985.
68. Fish, Stanley Eugene. *There's No Such Thing as Free Speech: And It's a Good Thing, Too.* Oxford: Oxford University Press, 1994.
69. Ford, Thomas E., Christie F. Boxer, Jacob Armstrong, and Jessica R. Edel. “More Than ‘Just a Joke’: The Prejudice-Releasing Function of Sexist Humor.” *Personality and Social Psychology Bulletin* 34, no. 2 (2008): 159–170. https://doi.org/10.1177/0146167207310022.
70. Fricker, Miranda. *Epistemic Injustice: Power and the Ethics of Knowing.* Oxford: Oxford University Press, 2007.
71. Fumagalli, Corrado. “Propositional Attitudes, Harm and Public Hate Speech Situations: Towards a Maieutic Approach.” *European Journal of Political Theory* 20, no. 4 (2021): 609-630. https://doi.org/10.1177/1474885119836627. Retrieved from https://philpapers.org/archive/FUMPAH-2.pdf.

72. Fuš, Mirela. “Pejoratives as Social Kinds: Objections to Miščević’s Account.” In *A Word Which Bears a Sword: Inquiries into Pejoratives,* edited by Nenad Miščević and Julija Perhat, 159-182. Zagreb: Kruzak, 2016.
73. Gard, Stephen W. “Fighting Words as Free Speech.” *Washington University Law Quarterly* 58 (1980): 531.
74. Gelber, Katharine. “Reconceptualizing Counterspeech in Hate-Speech Policy (With a Focus on Australia).” In *The Content and Context of Hate Speech,* edited by Michael Herz and Peter Molnar, 198-216. Cambridge: Cambridge University Press, 2012.
75. Goff, Phillip Atiba, Jennifer L. Eberhardt, Melissa J. Williams, and Matthew Christian Jackson. “Not Yet Human: Implicit Knowledge, Historical Dehumanization, and Contemporary Consequences.” *Journal of Personality and Social Psychology* 94, no. 2 (2008): 292–306. https://doi.org/10.1037/0022-3514.94.2.292.
76. Goguen, Stacey. “Stereotype Threat, Epistemic Injustice, and Rationality.” In *Implicit Bias and Philosophy Volume 1: Metaphysics and Epistemology,* edited by Michael Brownstein and Jennifer Saul, 147–172. Oxford: Oxford University Press, 2016.
77. Goldberg, Sandy. *Conversational Pressure.* Oxford: Oxford University Press, 2020.
78. Goodman, Jeffrey A., Jonathan Schell, Michele G. Alexander, and Scott Eidelman. “The Impact of a Derogatory Remark on Prejudice Toward a Gay Male Leader.” *Journal of Applied Social Psychology* 38, no. 2 (2008): 542–555. https://doi.org/10.1111/j.1559-1816.2008.00316.x.
79. Goodman, Mary Ellen. *Race Awareness in Young Children.* New York: Collier Macmillan, 1952.
80. Hailiang, Ning, Xue Dai, and Fachun Zhang. “On Gender Difference in English Language and Its Causes.” *Asian Social Science* 6, no. 2 (2010): 10.5539/ass.v6n2p126. Retrieved from http://www.ccsenet.org/journal/index.php/ass/article/view/5053/4201.
81. Hamilton, David L., and Robert K. Gifford. “Illusory Correlation in Interpersonal Perception.” *Journal of Experimental Social Psychology* 12 (1976): 392–407.

82. Harris, Monica J., Richard Milich, Elizabeth M. Corbitt, Daniel W. Hoover, and M. Brady. “Self-Fulfilling Prophecy Effects of Stigmatizing Information on Children’s Social Interactions.” *Journal of Personality and Social Psychology* 63 (1992): 41–50.
83. Haslanger, Sally. *Resisting Reality: Social Construction and Social Critique*. Oxford: Oxford University Press, 2012.
84. Haslanger, Sally. “Objectivity, Epistemic Objectification, and Oppression.” In *The Routledge Handbook of Epistemic Injustice,* edited by I. J. Kidd, J. Medina, and G. Pohlhaus Jr., 279–290. New York: Routledge, 2017.
85. Heider, Jeremy D., Cory R. Scherer, and John E. Edlund. “Cultural Stereotypes and Personal Beliefs About Individuals with Dwarfism.” *Journal of Social Psychology* 153, no. 1 (2013): 80-97. https://doi.org/10.1080/00224545.2012.711379.
86. Heinze, Eric. *Hate Speech and Democratic Citizenship*. Oxford University Press, 2016.
87. Hom, Christopher. “The Semantics of Racial Epithets.” *Journal of Philosophy* 105 (2008): 416–440.
88. Hom, Christopher. “Pejoratives.” *Philosophy Compass* 5 (2010): 164–185.
89. Hom, Christopher, and Robert May. “Pejoratives as Fiction.” In *Bad Words: Philosophical Perspectives on Slurs,* edited by David Sosa, 92–116. Oxford, United Kingdom: Oxford University Press, 2018.
90. Hornsby, Jennifer. “Illocution and Its Significance.” In *Foundations of Speech Act Theory: Philosophical and Linguistic Perspectives,* edited by Savas L. Tsohatzidis, 187–207. London: Routledge, 1994.
91. Hornsby, Jennifer, and Rae Langton. “Free Speech and Illocution.” *Legal Theory* 4 (1998): 21–37.
92. Hornsby, Jennifer. “Meaning and Uselessness: How to Think About Derogatory Words.” *Midwest Studies in Philosophy* 25, no. 1 (2001): 128–141.
93. Howard, Jeffrey W. “Free Speech and Hate Speech.” *Annual Review of Political Science* 22 (2019): X–X. https://doi.org/10.1146/annurev-polisci-051517-012343.
94. Howard, Jeffrey W. “Terror, Hate, and the Demands of Counterspeech.” *British Journal of Political Science* 51, no. 3 (2021): 924-939.

95. Howard, Jeffrey W. “Freedom of Speech.” In *The Stanford Encyclopedia of Philosophy* (Spring 2024 Edition), edited by Edward N. Zalta and Uri Nodelman. https://plato.stanford.edu/archives/spr2024/entries/freedom-speech/.
96. Hughes, Geoffrey. *An Encyclopedia of Swearing: The Social History of Oaths, Profanity, Foul Language, and Ethnic Slurs in the English-Speaking World.* New York and London: M.E. Sharpe, 2006.
97. International Convention on the Elimination of All Forms of Racial Discrimination. 1965.
98. International Covenant on Civil and Political Rights. *Article 19.* 1966.
99. Jeshion, Robin. “Expressivism and the Offensiveness of Slurs.” *Philosophical Perspectives* 27 (2013a): 315-329.
100. Jeshion, Robin. “Slurs and Stereotypes.” *Analytic Philosophy* 54, no. 3 (2013b): 315-329.
101. Jeshion, Robin. “Pride and Prejudiced.” *Grazer Philosophische Studien* 97, no. 1 (2020): 106-137.
102. Katz, Daniel, and K. W. Braly. “Racial Stereotypes of One Hundred College Students.” *Journal of Abnormal and Social Psychology* 28 (1933): 280–290.
103. Katz, Phyllis A. “Developmental Foundations of Gender and Racial Attitudes.” In *The Child’s Construction of Social Inequality,* edited by R. Leahy, 41–77. New York: Academic Press, 1983.
104. Keller, Johannes, and Dirk Dauenheimer. “Stereotype Threat in the Classroom: Dejection Mediates the Disrupting Threat Effect on Women’s Math Performance.” *Personality and Social Psychology Bulletin* 29, no. 3 (2003): 371–381.
105. Kelly, David J., Shaoying Liu, Liezhong Ge, Paul C. Quinn, Alan M. Slater, Kang Lee, Qinyao Liu, and Olivier Pascalis. “Cross-Race Preferences for Same-Race Faces Extends Beyond the African Versus Caucasian Contrast in 3-Month-Old Infants.” *Infancy* 11 (2007): 87–95.
106. Kukla, Rebecca. “Performative Force, Convention, and Discursive Injustice.” *Hypatia* 29, no. 2 (2014): 440–457. https://doi.org/10.1111/j.1527-2001.2012.01316.x.

107. Kulenović, Enes. “Should Democracies Ban Hate Speech? Hate Speech Laws and Counterspeech.” *Ethical Theory Moral Practice* 26 (2023): 511–532.

108. Lakoff, Robin. “Language and Woman’s Place.” *Language in Society* 2, no. 1 (1973): 45-80. Retrieved from: http://www.stanford.edu/class/linguist156/Lakoff_1973.pdf.

109. Langton, Rae. “Speech Acts and Unspeakable Acts.” *Philosophy & Public Affairs* 22, no. 4 (1993): 293–330.

110. Langton, Rae. “The Authority of Hate Speech.” In *Oxford Studies in Philosophy of Law, Vol. 3,* edited by John Gardner, Leslie Green, and Brian Leiter, 123–152. Oxford: Oxford University Press, 2018.

111. Langton, Rae. “Blocking as Counterspeech.” In *New Work on Speech Acts,* edited by Daniel Fogal, Daniel W. Harris, and Matt Moss, 144–162. Oxford: Oxford University Press, 2018.

112. Law on Misdemeanors Against Public Order and Safety. 2023. *Official Gazette* 41/77, 52/87, 47/89, 55/89, 5/90 (consolidated text), 30/90 (correction of consolidated text), 47/90, 29/94, and 114/22. Croatian Parliament. Published on May 3, 2023.

113. Lawrence, Charles III. “If He Hollers Let Him Go: Regulating Racist Speech on Campus.” In *Words That Wound: Critical Race Theory, Assaultive Speech, and the First Amendment*, edited by Mari Matsuda, Charles Lawrence III, Richard Delgado, and Kimberlé Crenshaw, 53–88. Boulder, CO: Westview Press, 1993.

114. Lederer, Laura, and Richard Delgado. “Preface to the Book.” In *The Price We Pay: The Case Against Racist Speech, Hate Propaganda, and Pornography,* New York: Hill and Wang, 1995.

115. Leader Maynard, Jonathan, and Susan Benesch. “Dangerous Speech and Dangerous Ideology: An Integrated Model for Monitoring and Prevention.” *Genocide Studies and Prevention: An International Journal* 9, no. 3 (2016): 70-95.

116. Legaspe, Justina Diaz**.** “Normalizing Slurs and Out-Group Slurs: The Case of Referential Restriction.” *Analytic Philosophy* 59, no. 2 (June 2018): 1–22. https://philpapers.org/archive/LEGNSA.pdf.

117. Lepoutre, Maxime. “Hate Speech in Public Discourse: A Pessimistic Defense of Counterspeech.” *Social Theory and Practice* 43, no. 4 (2017): 851–883.

118. Levy, Neil. “No-Platforming and Higher-Order Evidence or Anti-Anti-No Platforming.” *Journal of the American Philosophical Association* 5, no. 4 (2019): 487-502.

119. Lewandowsky, Stephan, Ulrich Ecker, Colleen Seifert, Norbert Schwarz, and John Cook. “Misinformation and Its Correction: Continued Influence and Successful Debiasing.” *Psychological Science in the Public Interest* 13, no. 3 (2012): 106–131.

120. Lippmann, Walter. *Public Opinion.* New York: Free Press, 1965.

121. Liu, Chang. “The Derogatory Force and the Offensiveness of Slurs.” *Organon F: Medzinárodný Časopis Pre Analytickú Filozofiu* 28, no. 3 (2021): 626–649.

122. Maass, Anne, Daniela Salvi, Luciano Arcuri, and Gün R. Semin. “Language Use in Intergroup Contexts: The Linguistic Intergroup Bias.” *Journal of Personality and Social Psychology* 57 (1989): 981–993.

123. Maccoby, Eleanor E., and Carol Nagy Jacklin. “Gender Segregation in Childhood.” *Advances in Child Development and Behaviour* 20 (1987): 239–287.

124. Maccoby, Eleanore E. “Historical Overview of Socialization Research and Theory.” In *Handbook of Socialization,* edited by JE Grusec and PD Hastings, 13–41. New York: The Guilford Press, 2007.

125. Mackenzie, Catriona, and Denise Meyerson. “Autonomy and Free Speech.” In *The Oxford Handbook of Freedom of Speech,* edited by Adrienne Stone and Frederick Schauer. Oxford: Oxford University Press, 2021. Online edition.

126. MacKinnon, Catharine. *Only Words.* Cambridge, MA: Harvard University Press, 1993.

127. Madon, Stephanie, Alison Smith, Lee Jussim, Daniel Russell, Jacquelynne S. Eccles, Michele Walkiewicz, and Polly Palumbo. “Am I as You See Me or Do You See Me as I Am? Self-Fulfilling Prophecies and Self-Verification.” *Personality and Social Psychology Bulletin* 27 (2001): 1214–1224.

128. Maitra, Ishani. “Silencing Speech.” *Canadian Journal of Philosophy* 39, no. 2 (2009): 309–338.

129. Marques, Teresa. "'Beasts in Human Form': How Dangerous Speech Harms." *Araucaria. Revista Iberoamericana de Filosofía, Política y Humanidades* 21, no. 42 (2019): 553-584.

130. Matsuda, Mari J. "Public Response to Racist Speech: Considering the Victim's Story." *Michigan Law Review* 87, no. 8 (1989): 2320–2381.

131. McGowan, Mary Kate. "Conversational Exercitives: Something Else We Do with Our Words." *Linguistics and Philosophy* 27 (2004): 93–111.

132. McGowan, Mary Kate. "Oppressive Speech." *Australasian Journal of Philosophy* 87, no. 3 (2009): 389–407.

133. McGowan, Mary Kate. "Sincerity Silencing." *Hypatia* 29, no. 2 (2014): 458–473.

134. Mead, George Herbert. *Mind, Self, and Society.* Edited by C. W. Morris. Chicago: University of Chicago Press, 1962.

135. Media Act. 2013. *Official Gazette* 59/2004, 84/2011, 81/2013. Croatian Parliament.

136. Medina, José. *The Epistemology of Resistance: Gender and Racial Oppression, Epistemic Injustice, and the Social Imagination.* Oxford University Press, 2013.

137. Meiklejohn, Alexander. *Free Speech and Its Relation to Self-Government.* New York: Harper, 1948.

138. Mill, John Stuart. *On Liberty and The Subjection of Women.* New York: Henry Holt and Co., 1879. https://oll-resources.s3.us-east-2.amazonaws.com/oll3/store/titles/347/Mill_0159_EBk_v6.0.pdf.

139. Mikkola, Mari. "Illocution, Silencing and the Act of Refusal." *Pacific Philosophy Quarterly* 92 (2011): 415–437.

140. Mikkola, Mari. *Pornography: A Philosophical Introduction.* New York: Oxford University Press, 2019.

141. Miščević, Nenad. "The Fiery Tongue: Semantics and Pragmatics of Pejoratives." In *A Word Which Bears a Sword: Inquiries into Pejoratives*, edited by Nenad Miščević and Julija Perhat, 15–158. Zagreb: Kruzak, 2016.

142. Miščević, Nenad, and Julija Perhat, eds. *A Word Which Bears a Sword: Inquiries into Pejoratives.* Zagreb: Kruzak, 2016.

143. Moles, Andres. “Autonomy, Free Speech and Automatic Behaviour.” *Res Publica* 13 (2007): 53-75.

144. Nagel, Thomas. *Concealment and Exposure.* New York: Oxford University Press, 2002.

145. Nesdale, Drew. “Social Identity Processes and Children’s Ethnic Prejudice.” In *The Development of the Social Self,* edited by Mark Bennett and Felicity Sani, 219–245. Hove: Psychology Press, 2004.

146. Nunberg, Geoff. “The Social Life of Slurs.” In *New Work on Speech Acts,* edited by Daniel Fogal, Daniel W. Harris, and Matt Moss, 237–295. Oxford: Oxford University Press, 2018.

147. Oxford Advanced Learner’s Dictionary. 8th ed. Oxford University Press, 2010.

148. Parsons, Jacquelynne Eccles, Terry F. Adler, and Caroline M. Kaczala. “Socialization of Achievement Attitudes and Beliefs: Parental Influences.” *Child Development* 53 (1982): 310–21.

149. Perhat, Julija. “A Word Which Bears a Sword: The Semantic and Ethico-Political Dimensions of Gender Pejoratives.” Unpublished MA thesis, 2012.

150. Perhat, Julija. “Pejoratives and Testimonial Injustice.” In *A Word Which Bears a Sword: Inquiries into Pejoratives*, edited by Nenad Miščević and Julija Perhat, 123–145. Zagreb: Kruzak, 2016.

151. Popa-Wyatt, Mihaela, and Jeremy L. Wyatt. “Slurs, Roles and Power.” *Philosophical Studies* 175 (2018): 2879–2906. https://doi.org/10.1007/s11098-017-0986-2.

152. Post, Robert. “Hate Speech.” In *Extreme Speech and Democracy,* edited by Ivan Hare and James Weinstein, Oxford: Oxford University Press, 2009. Online edition.

153. Potts, Christopher. “The Expressive Dimension.” *Theoretical Linguistics* 33, no. 2 (2007): 165–197.

154. Prijić-Samaržija, Snježana. “Epistemic Virtues of Institutions.” In *Institutions in Action: The Nature and the Role of Institutions in the Real World,* edited by T. Andina and P. Bojanić, 21-36. Cham: Springer, 2020.

155. Samaržija, Hana, and Ivan Cerovac. “The Institutional Preconditions of Epistemic Justice.” *Social Epistemology* published online 2021. https://doi.org/10.1080/02691728.2021.1919238.

156. Rawls, John. *A Theory of Justice: Original Edition.* Cambridge, MA: Harvard University Press, 1971. https://doi.org/10.2307/j.ctvjf9z6v.

157. Reid, Andrew. “Does Regulating Hate Speech Undermine Democratic Legitimacy? A Cautious ‘No’.” *Res Publica* 26, no. 2 (2020): 181-199.

158. Richard, Mark. *When Truth Gives Out.* Oxford: Oxford University Press, 2008.

159. Riesman, David. “Democracy and Defamation: Control of Group Libel.” *Columbia Law Review* 42 (1942): 727–780.

160. Rosenthal, Robert, and Lenore Jacobson. *Pygmalion in the Classroom: Teacher Expectations and Student Intellectual Development.* New York: Holt, Rinehart, and Winston, 1968.

161. Rosette, Ashleigh Shelby, Andrew M. Carton, Lynn Bowes-Sperry, Patricia Faison Hewlin. “Why Do Racial Slurs Remain Prevalent in the Workplace? Integrating Theory on Intergroup Behavior.” *Organization Science* 24, no. 5 (2013): 1402–1421. https://doi.org/10.1287/orsc.1120.0809.

162. Samson, Edward E. *Dealing with Differences: An Introduction to the Social Psychology of Prejudice.* New York: Harcourt Brace, 1999.

163. Saul, Jennifer. “Implicit Bias, Stereotype Threat, and Women in Philosophy.” In *Women in Philosophy: What Needs to Change,* 39-60. New York: Oxford University Press, 2013.

164. Saul, Jennifer. “Implicit Bias, Stereotype Threat, and Epistemic Injustice.” In *The Routledge Handbook of Epistemic Injustice,* edited by I.J. Kidd, J. Medina, and G. Pohlhaus, 235-243. London: Routledge, 2017.

165. Saul, Jennifer. “Someone Is Wrong on the Internet: Is There an Obligation to Correct False and Oppressive Speech on Social Media?” In *The Epistemology of Deceit in a Postdigital Era,* edited by Alison MacKenzie, Jennifer Rose, and Ibrar Bhatt, 139–157. New York: Springer, 2021.

166. Saunders, W. Kevin. *Degradation: What the History of Obscenity Tells Us about Hate Speech.* New York and London: New York University Press, 2011.

167. Scanlon, Thomas. "A Theory of Freedom of Expression." *Philosophy & Public Affairs* 1, no. 2 (1972): 204–226.

168. Scanlon, Thomas. *The Difficulty of Tolerance: Essays in Political Philosophy.* Cambridge: Cambridge University Press, 2003.

169. Seglow, Jonathan. "Hate Speech, Dignity and Self-Respect." *Ethical Theory and Moral Practice* 19 (2016): 1103–1116. https://doi.org/10.1007/s10677-016-9744-3.

170. Semin, Gün R., and Klaus Fiedler. "The Cognitive Functions of Linguistic Categories in Describing Persons: Social Cognition and Language." *Journal of Personality and Social Psychology* 54 (1988): 558–568.

171. Shapiro, Jenessa, and Joshua Aronson. "Stereotype Threat." In *Stereotyping and Prejudice,* edited by Charles Stangor and Christian Crandall, 95–117. New York: Psychology Press, 2013.

172. Shiffrin, Seana Valentine. *Speech Matters: On Lying, Morality, and the Law*. Princeton: Princeton University Press, 2014. http://www.jstor.org/stable/j.ctt9qh09j.

173. Shklar, Judith. *The Faces of Injustice.* New Haven and London: Yale University Press, 1990.

174. Simpson, Robert Mark. "Dignity, Harm, and Hate Speech." *Law and Philosophy* 32, no. 6 (2013): 701-728.

175. Simpson, Robert Mark. "'Won't Somebody Please Think of the Children?' Hate Speech, Harm, and Childhood." *Law and Philosophy* 38, no. 1 (2019): 79–108.

176. Soral, Wiktor, Michał Bilewicz, and Mikołaj Winiewski. "Exposure to Hate Speech Increases Prejudice Through Desensitization." *Aggressive Behavior* 44, no. 2 (2018): 136–146. https://doi.org/10.1002/ab.21737.

177. Stangor, Charles, Christine Can, and Lisa Kiang. "Activating Stereotypes Undermines Task Performance Expectations." *Journal of Personality and Social Psychology* 75 (1998): 1191-1197.

178. Stangor, Charles, and Christian Crandall, eds. *Stereotyping and Prejudice.* New York: Psychology Press, 2013.

179. Steele, Claude M., and Joshua Aronson. “Stereotype Threat and the Intellectual Test Performance of African Americans.” *Journal of Personality and Social Psychology* 69, no. 5 (1995): 797–811.

180. Stojnić, Una, and Ernie Lepore. *Inflammatory Language: Its Linguistics and Philosophy.* Forthcoming. Oxford: Oxford University Press.

181. Swim, Janet K., Kristen Johnston, and Nicholas B. Pearson. “Daily Experiences with Heterosexism: Relations Between Heterosexist Hassles and Psychological Well-Being.” *Journal of Social and Clinical Psychology* 28, no. 5 (2009): 597–629.

182. Tsesis, Alexander. *Destructive Messages: How Hate Speech Paves the Way for Harmful Social Movements.* New York: NYU Press, 2002.

183. United Nations. *Universal Declaration of Human Rights.* 1948.

184. United Nations. *Strategy and Plan of Action on Hate Speech.* 2019. https://www.un.org/en/genocideprevention/documents/advising-and-mobilizing/Action_plan_on_hate_speech_EN.pdf.

185. Volpato, Chiara, Federica Durante, Alessandro Gabbiadini, Luca Andrighetto, and Silvia Mari. “Picturing the Other: Targets of Delegitimization across Time.” *International Journal of Conflict and Violence* 4, no. 2 (2010): 269-287. https://doi.org/10.4119/ijcv-2831.

186. Waldron, Jeremy. *The Harm in Hate Speech.* Cambridge, MA: Harvard University Press, 2012.

187. Walker, Samuel. *Hate Speech: The History of an American Controversy.* Lincoln: University of Nebraska Press, 1994.

188. Weinstein, James. “Hate Speech Bans, Democracy, and Political Legitimacy.” *Constitutional Commentary* 32 (2017): 527–583.

189. Williamson, Timothy. “Reference, Inference, and the Semantics of Pejoratives.” In *The Philosophy of David Kaplan,* edited by J. Almog and P. Leonardi, 137–159. Oxford University Press, 2009.

190. Wilson, Timothy D., and Nancy Brekke. “Mental Contamination and Mental Correction: Unwanted Influences on Judgments and Evaluations.” *Psychological Bulletin* 116 (1994): 117-142.

191. Winiewski, Mikołaj, Karolina Hansen, Wiktor Soral, Aleksandra Świderska, Dominika Bulska, and Michał Bilewicz. *Contempt Speech, Hate Speech. Report from Research on Verbal Violence Against Minority Groups.* Stefan Batory Foundation, 2017. http://www.ngofund.org.pl/wp-content/uploads/2017/02/Contempt_Speech_Hate_Speech_Full_Report.pdf.

192. Yee, Mia D., and Rupert Brown. *Children and Social Comparisons.* Swindon: Economic and Social Research Council, 1988.

193. Yule, George. *The Study of Language.* Cambridge: Cambridge University Press, 1996.

194. Zeman, Dan. "A Rich-Lexicon Theory of Slurs and Their Uses." *Inquiry* 65, no. 7 (2021): 942–966. https://doi.org/10.1080/0020174X.2021.1903552.

LIST OF TABLES

LIST OF FIGURES