TOWARDS A BETTER UNDERSTANDING OF MORALLY RESPONSIBLE AGENCY

A Master's Thesis

by

Roohollah Haghshenas

Department of Philosophy

İhsan Doğramacı Bilkent University

Ankara

May 2021

To

Superbad (and Supergood) People This Thesis Calls for Understanding Them

TOWARDS A BETTER UNDERSTANDING OF MORALLY RESPONSIBLE AGENCY

The Graduate School of Economics and Social Sciences
of

İhsan Doğramacı Bilkent University

by

Roohollah Haghshenas

In Partial Fulfillment of the Requirements for the Degree of

MASTER OF PHILOSOPHY

THE DEPARTMENT OF PHILOSOPHY
İHSAN DOĞRAMACI BİLKENT UNIVERSITY

ANKARA

May 2021

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Philosophy.

Assoc. Prof William Giles Wringe
Supervisor

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Philosophy.

Prof Neil Levy (Oxford University/Macquarie University)
Examining Committee Member

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Philosophy.

Asst. Prof. Jonathan Payton (Bilkent University)
Examining Committee Member

Approval of the Graduate School of Economics and Social Sciences

---

Prof Refet Soykan Gürkaynak
Director

ABSTRACT

TOWARDS A BETTER UNDERSTANDING OF MORALLY RESPONSIBLE AGENCY

Haghshenas, Roohollah
MA, Department of Philosophy
Supervisor: Prof. Dr. Bill Wringe

May 2021

In this thesis, I defend P. F. Strawson's distinction of internal-external problems to our ideas of moral responsibility practices. Then, I introduce the problems of superbad people as some serious internal problems. What I call Moral Personality Disorders, like narcissism, and deep-seated racism can be some instances of being superbad. I argue that just being superbad may make blame unintelligible for the blamed person, may make reactions like sadness appropriate to him, and may make blame an obstacle to finding deep roots of his problem and some effective solutions for it. I conclude that these problems prove the need for some substantial modifications in our ideas of moral responsibility. I ground a new account of responsibility based on what I introduce as one's quality of valuing (QV) and a historical condition of responsibility. The historical condition, I argue, is met through a Responsibility Chain: 1) We are responsible for our actions/choices as much as they are up to our QV at the time of doing them. 2) We are responsible for our QV at any given time as much as it is up to our previous actions/choices. As its negative force, the Responsibility Chain shows that the credit and discredit of our actions/choices cannot go to a self over than, and beyond to, the Responsibility Chain of our lives. The Responsibility Chain also shows why superbad (and supergood) people are some natural results of human nature and how we should react to them.

Keywords: Moral Responsibility, Blame, Personality Disorders, Quality of Will, Valuing

ÖZET

Ahlaki Sorumlu Kurumun Daha İyi Anlaşılmasına

Haghshenas, Roohollah

MA, Felsefe Bölümü

Tez Danışmanı: Prof. Dr. Bill Wringe

Mayıs 2021

Bu tezde, P.F. Strawson'un iç-dış problemler ayrımını ahlaki sorumluluk uygulamaları fikirlerimizle savunuyorum. Sonra, süper kötü insanların sorunlarını bazı ciddi iç sorunlar olarak tanıtıyorum. Narsisizm ve köklü ırkçılık gibi Ahlaki Kişilik Bozuklukları dediğim şey, süper kötü olmanın bazı örnekleri olabilir. Sadece süper kötü olmanın, suçlanan kişiyi suçu anlaşılmaz hale getirebileceğini, üzüntü gibi tepkiler verebileceğini ve sorunun derin köklerini ve bazı etkili çözümlerini bulmaya engel teşkil edebileceğini savunuyorum. Bu sorunların ahlaki sorumluluk fikirlerimizde bazı önemli değişikliklere olan ihtiyacı kanıtladığı sonucuna vardım. Kişinin değer verme kalitesi (QV) ve tarihsel bir sorumluluk durumu olarak sunduğum şeye dayalı olarak yeni bir sorumluluk hesabı oluşturdum. Tarihsel koşulun bir Sorumluluk Zinciri aracılığıyla karşılandığını ileri sürüyorum: 1) Eylemlerimizden / seçimlerimizden, bunları yaparken QV'mize bağlı olduğu kadar sorumluyuz. 2) Önceki eylemlerimize / seçimlerimize bağlı olduğu kadar, herhangi bir zamanda QV'mizden sorumluyuz. Negatif gücü olarak Sorumluluk Zinciri, eylemlerimizin / seçimlerimizin itibarının ve itibarsızlığının, hayatımızın Sorumluluk Zincirinin ötesinde ve ötesine geçemeyeceğini göstermektedir. Sorumluluk Zinciri ayrıca süper kötü (ve süper iyi) insanların neden insan doğasının bazı doğal sonuçları olduğunu ve onlara nasıl tepki vermemiz gerektiğini gösterir.

Anahtar Kelimeler: Ahlaki Sorumluluk, Suçlama, Kişilik Bozuklukları, İrade Kalitesi, Değerleme

ACKNOWLEDGMENTS

It was not easy for me to find how to formulate the concerns I develop in this thesis and to find the literature they can be related to. My first and deepest thanks go to those who helped me in the quite long years behind starting this thesis and even my studies at Bilkent: Mohammad Sadegh Zahedi, Amir Saemi, Jeanette Kennett, Philip Pettit, and the extreme generosity of Victoria McGeer and Neil Levy. Abolfazl Sabramiz is the one who I cannot imagine how I could survive those years without his help. I also never forget the encouragement and support of my mother, Khosrow Ghobadi, Shahin Akhondzadeh, Mahdi Zakeri, Hadi Khaniki, and most notably, my sister, Samaneh. In the phase of working on my ideas more professionally, I am indebted to the many generous comments of Hannah Tierney, Angela Smith, Victoria McGeer, Jonathan Payton, and again, Neil Levy. How wonderful Bill Wringe is as a supervisor and mentor is the most frequent topic of my discussions with other graduate students at Bilkent. I am happy to have the opportunity of knowing these wonderful people in my life.

TABLE OF CONTENTS

INTRODUCTION

A few years before reading any philosophical work about moral responsibility, I and some friends were following closely the extremely violent responses of a dictatorship to an entirely non-violent protest movement. The way the people involved in the movement put their message to their opponents was so beautiful and touching that I couldn't understand how the regime's proponents, even the dictator himself, could resist it, kill people horribly, and rape them in prisons. I remember I told one of my friends that I didn't know whether anger or sadness is the appropriate reaction to the dictator and his proponents. My friend got angry with me and asked, "What else do they need to do to deserve anger?" He was right. There was no doubt about the abhorrence of their actions. Still, I asked my friend, "Suppose you could choose to have all they have—political power, secure jobs, etc.—just if you accept to be a person like them. Would you choose it?" He answered no, without hesitation. I told him, "This alone might be a good reason for sadness." On the other hand, the people I was concerned about were not blind forces of nature, nor were the conditions philosophers suggest as exempting—severe mental illnesses, nonculpable ignorance, extremely misfortunate upbringing, and so on—true of most of them. In this thesis, I try to reveal the problems like this that are, in Strawson's terms, internal to our practices of responsibility. To remedy these problems, I suggest a quality-of-valuing account of moral responsibility that I hope better describes what kind of morally responsible agents we are.

CHAPTER 1:

DESERVED ANGER AND THE STRENGTH OF OUR MORAL RESPONSIBILITY PRACTICES

Introduction

Moral responsibility, or morally responsible agency, is a philosophical concept that on one hand seems to constitute some important part of our personal and inter-personal lives and on the other hand skepticism about it, mostly from the worry about its compatibility with determinism, has been in the center of long-lasting philosophical debates. To have a clearer idea of these two sides, we need a clearer idea of each.

I think moral responsibility practices—hereafter, MRP—are based on our having a quite clear idea of 1) When we should and when we should not hold people morally responsible for their conduct—that is the idea of *moral responsibility judgment* and the necessary and sufficient conditions of it, and 2) How we should react to those who we come to hold them morally responsible for their conduct—that is the idea of *moral responsibility reactions*. Philosophical skepticism about moral responsibility apart for a moment, most people have some answer, even vague, to these two questions. We think differently about a person who uses seemingly offensive words against us when we come to know that she does *not* know that the words are offensive or that she had gone through some hard times recently. Following these changes in our judgment of her, we react to her accordingly. Our judgment

of people's undesirable actions can change because of more fundamental facts about them. This is why we know that children or people with some abnormal cognitive abilities should be treated differently. There are many other things like these that might show that we have some relatively clear ideas of moral responsibility judgment and moral responsibility reactions.

Skepticism about moral responsibility, in most common versions, is the claim that at least in some important sense of moral responsibility, there is no morally responsible agent either in our actual world because, say, this world is deterministic, or in any possible world because of some self-inconsistencies.

MRP—that is our idea of moral responsibility judgments and our idea of moral responsibility reactions—are not just some superstitious beliefs that every—or even most—knowledgeable people know are false. Based on this prima facie strength, P. F. Strawson grounds his argument against philosophical debates on the determinism-based skepticism about moral responsibility because each side, he thinks, "over-intellectualize the facts". In this chapter, I examine Strawson's account and some criticisms of it. I distinguish Strawson's "picture" of our MRP and his "account" according to which the picture proves the skepticism about moral responsibility unimportant. I defend Strawson's "picture" as a worthy picture of our MRP. About Strawson's "account", I suggest that both he and his opponent have the not-so-easy task of defining the desert that the reactive attitudes seem to presuppose to prove whether or not determinism is an important worry. For the purpose of my following chapters, I defend a weaker claim than Strawson's "account", namely, the claim that the problems *internal* for our MRP—based on Strawson's "picture"—should be prioritized to the problems like the controversial threat of determinism. Remaining faithful to this weaker claim, I will discuss the accounts of T. M Scanlon, Nomy Arpaly, and Angela Smith in chapters 3 and 4 as the more recent versions of a quality of will accounts to ground my positive solution to some "internal" problems of responsibility I introduce in chapter 2.

I. Strawson's three theses against skepticism

P. F. Strawson's "Freedom and Resentment" (1982 [1962]) is one of the most influential works on the last six decades of thinking to moral responsibility. In this section, I try to uncover Strawson's three theses in his paper against determinism-based skepticism about moral responsibility. In short, I call Strawson's theses the Not-Relevant thesis; Naturalist thesis; and Practical Rationality thesis.

Before continuing, let us see how Strawson sees the debate existing at his time. First, there is a version of compatibilism Strawson calls Optimism: 1) blame and praise are influential on the future behaviors of their targets, 2) to blame and praise a person is to hold her morally responsible, and 3) this notion of responsibility is compatible with the truth of determinism. Those who Strawson calls Pessimists, on the other hand, see a contra-causal notion of self necessary for responsibility that makes them believe in the incompatibility of responsibility and determinism.

Strawson's argument against Optimists is straightforward. Blame and praise are influential also on the future behaviors of animals and young children. But as animals and children are not (fully) responsible, the optimists' notion of responsibility is too broad. Strawson shares the Optimists' goal in showing the compatibility of responsibility and determinism but by grounding it on a better picture of our MRP. Strawson's main picture is that the whole idea of moral responsibility is: I) to see when one's behavior shows her ill or good will, and II) when I is the case, to react to her with reactive attitudes and feelings like resentment or indignation—or guilt in the case of self-reflection—to the agent's ill will or with gratitude to her goodwill.

Strawson expands I by arguing that if we set aside the cases in which one has a *justification*, like when one passes the red light because there is a person with an emergency medical condition in her car, or an *excuse*, like when one drives on the wrong side of the road because she has come to the country recently and doesn't know the difference with her own country, or when one is *exempted*, because of things like being too young or some cognitive disabilities, all other cases of wrong conduct show one's ill will, like acting because of being too selfish to care about others' rights.

It is important to note that being exempted in Strawson's view is not something that a normal person has reason to want—unlike being exempted from some legal punishments. Reacting to a wrongdoer, in Strawson's view, means holding her a competent member of the moral community—this is what Strawson calls, "the participant stance"—while being exempted means being a case that needs to be "managed or handled or cured or trained", that is to be treated as *it*, either temporarily or permanently—this is what Strawson calls, "the objective stance." Having this general picture of Strawson's project, I explain Strawson's three theses.

First, Strawson's *Not-Relevant thesis* is a thesis that as determinism is not a thesis that all people should be held exempted and treated only from the objective stance—and not, of course, a thesis that all people have always a justification or excuse for their wrongdoings—so, the truth or falsity of determinism is not relevant to the justification of our MRP.

Second, Strawson's *Naturalist* thesis is that I and II are too deep in our nature that even if some philosophical arguments show that determinism implies that both exempted and not-exempted cases should be treated in the same way—in term of what is significant from a moral responsibility point of view—we cannot make this change in our psychology. Strawson's naturalism is a Humean naturalism according to which the reactive attitudes, like any other type of our emotions, are open to case-by-case modifications but not to any "external justification": we just have them and for this, we need no rational justification.

Finally, Strawson's Practical Rationality thesis is that even if determinism implies that both exempted and not-exempted cases should be treated in the same way *and* we can make this change in our psychology, we should compare what we gain and what we lose by such a change to see whether it is practically rational or not. Strawson's own answer to this comparison is that what we will lose by making this change, that is the many important personal and inter-personal emotions, is much more than what we gain, that is only being in more accordance with some the conclusion of some (if any) philosophical arguments.

I suggested the three theses above as distinct theses. I think there is no thesis in Strawson's paper beyond these three. But I see it possible to read Strawson in a way that only with a combination of all or two theses we can have a complete Strawsonian argument. I also did

not discuss Strawson's list of exempting (and justifying and excusing) conditions and whether his list is comprehensive enough. For my later discussions, both these two points do not matter so much.

II. Evaluating Strawson's move

As I said at the beginning of this chapter, I am trying to see how strong and robust our moral responsibility practice—or our ideas of moral responsibility judgments and moral responsibility reaction—are. I am studying Strawson as one of the best moves to approaching the answer to this question. After introducing Strawson's thesis, in this section. I try to evaluate how successful Strawson's move is. I put my evaluation of Strawson's move in two positive points in favor of and one negative point against him: 1. His claim that intuitively, there are some exempting conditions that can change the moral responsibility of people significantly is compelling. Consider kleptomania as a condition that makes a person steal something even when it is clearly against his best interests—for example, stealing a cheap pen of his boss when the risk of being caught and fired is too high just because the pen is shiny and it triggers one to steal it. 2. The exempted/responsible distinction points to something more substantive than the ability to act voluntarily and intentionally. Young children do many things voluntarily and intentionally but for which they are not fully responsible. 3. Although Strawson is successful in showing 2, his highest goal to show that the intuitive exempted/responsible distinction is substantive enough to rule out the determinism threat is at best an incomplete task. I hope my discussion in the previous section could have supported 1 and 2. In this section, I argue for 3.

Strawson's thesis that determinism is not relevant to the practices of holding people responsible has faced the objection that if determinism is true no one is deserving of the reactive attitudes like resentment or guilt. Resentment, the paradigm case of reactive attitudes, is a kind of anger that its target is held deserving of. This desert base of reactive attitudes is their difference with other instances of anger that we think we should avoid having and if we have on some occasions, we may come to regret it on giving it a second thought. To avoid merely verbal disputes, Derk Pereboom suggests that the sense of moral responsibility that is at issue in the debates of moral responsibility skepticism is the basic-

desert involving sense, that is when one is held deserving of blame or praise "just because she has performed the action, given an understanding of its moral status, and not, for example, merely by virtue of consequentialist or contractualist considerations" (Pereboom 2014: 2).

One influential challenge to Strawson's Not-Relevance thesis of determinism can be seen in Derk Pereboom's Manipulation Argument. Although it is an argument against any version of compatibilism, including Strawson's, reading it specifically against Strawson states that if someone's ill will against us is the result of some certain manipulation by a neurosurgeon on her mind shortly before her action or the result of programming her mind from the time of her birth to the time of her action, she is not deserving of our blame. Pereboom, then, challenges compatibilists to show a significant difference between a deterministic world and a world in which all agents are programmed throughout their lives. The challenge is if no one in the latter is deserving of blame and praise in the basic sense, how anyone can be deserving of blame and praise in the basic sense in the first. I do not discuss the literature on Pereboom's Manipulation Argument. My aim is just to show the centrality of basic desert in the debates on Strawson.

If reactive attitudes can lose their link to desert, Strawson's Naturalist thesis might become important—the idea that due to our psychological nature, it is impossible for us to leave the reactive attitudes in non-exempted cases. Against this thesis, Pereboom, says that in the case of any "unfair or irrational attitudes towards others that were difficult or even impossible to eradicate [… one should be at least] sad and embarrassed that one has the attitudes in question, avoids indulging or reveling in them, does what one can to rid oneself of them, and one certainly does not justify practical decisions on their basis (Pereboom, 2001: 98-99)." I think Pereboom is right and these things *are* possible for us and hence we can see whether we have reasons for trying them.

Against Strawson's claim that our system of reactive attitudes as a whole cannot be a subject of rational modification, Pereboom's points out that "analogies from other areas of ethical concern show that a system of attitudes [like some sexist and racist attitudes] can be subject to justificatory pressures from highly general theoretical beliefs (ibid, 98).

Strawson's thesis on the practical rationality of reactive attitudes has not been immune to criticism as well. This criticism maintains that some non-desert-based moral reactions like moral sadness, based on some kind of love to wrongdoers, can be good, or even better, replacements for Strawsonian reactive attitudes. I just mention two points proposed by Watson (2016).

As Watson quotes from Einstein, he thinks, following Schopenhauer, that "[a] man can do what he wants, but not want he wants". As Einstein concludes, "[t]his realization mercifully mitigates the easily paralyzing sense of responsibility and prevents us from taking ourselves and other people all too seriously (Watson, 2016)." Einstein's point was about the merits of withdrawing, or at least mitigating, holding people and ourselves responsible. Another point related to the practical rationality of abandoning desert-based—or in Watson terms, "retributive"—reactive attitudes can be seen in the views of people like Gandhi or Luther King. As Watson points out, they

> hold themselves and others morally responsible: They *stand up* for themselves and others against their oppressors; they *confront* their oppressors with the fact of their misconduct, *urging* and even *demanding* consideration for themselves and others; but they manage, or come much closer than others to managing, to do such things without vindictiveness or malice (Watson, 2016).

In chapter two, I back to the positive practical rationality of leaving the desert-based angry blame. Here, I just want to suggest that it is not as clear as Strawson thinks that the practical point of keeping the desert-based angry blame is much higher than the practical point of leaving it.

III. The desert base of reactive attitudes; a deeper look

Although Pereboom's clarification of what basic desert is is valuable, it is still not clear enough. I think we can differentiate a weak sense of basic desert from a strong one. The strong sense is what Galen Strawson describes as the target of his skepticism: "responsibility of such a kind that, if we have it, then it *makes sense*, at least, to suppose that it could be just to punish some of us with (eternal) torment in hell and reward others with (eternal) bliss in heaven (Strawson, 1994)".

But it is not clear that Strawsonian reactive attitudes necessarily presuppose such a strong sense of desert. Consider a view that holds your anger appropriate to someone who wronged you just because it is a necessary part of your valuing your relationship with her. Samuel Scheffler (2011) has argued that being susceptible to feeling some emotions related to something is a necessary part of valuing it. Suppose valuing your relationship with a sibling who lives far from you. To say you value this relationship, you should feel happy if you hear she is coming to your city and you should feel disappointed if you hear later that her trip is canceled. Drawn on Scheffler's idea, Agnes Callard has argued that you have reasons for being angry at someone who has wronged you only because his wrong has diminished the co-valuing relationship you have had with him. Suppose you hear that your sibling has come to your city without letting you know. Or suppose a harsh joke your partner makes about your physical attractiveness. Callard thinks that we are in a relationship of being members of a moral community with other humans and it gives us reasons to be angry at those who diminish this relationship. I will back to Callard in the following chapter. Here, I just want to suggest her view as a view that although is much weaker than Galen Strawson's sense, satisfies Pereboom's criterion basic desert—by claiming that you have reasons for anger at a wrongdoer "just because she has performed the action, given an understanding of its moral status, and not, for example, merely by virtue of consequentialist or contractualist considerations" (Pereboom 2014: 2). One may think that Callard's view is not a desert view at all because desert is a feature of only some sorts of punishment but when anger is not expressed it is not an instance of punishment. I disagree. Suppose I have been angry at you for a while and my anger has been ceased a while ago because I have found that it has been based on some wrong assumptions about you. I think when things like my relationship with you, the duration and severity of my anger, etc., are significant enough, you may object to me that I should have checked my assumption earlier (by letting you know about my anger or by any other ways) because you think you have been not deserving of my anger during that time.

Can we say whether Strawsonian reactive attitudes are based on a Galen Strawsonian strong sense of desert or Callardian weak sense? I think most people who blame Hitler and praise Gandhi think that "it *makes sense*, at least, to suppose that it could be just to punish [Hitler] with (eternal) torment in hell and reward [Gandhi] with (eternal) bliss in heaven."

It may suggest that the apparent difference in the strength of Galen Strawsonian and Callardian senses of desert is only because the first view is put in a way to be about extreme cases like Hitler and Gandhi but the latter is not. I disagree. The hell-heaven view can be read to be, also, about mundane right- or wrongdoers. In this reading, it states that every (small) wrongdoing can make you deserving of something a bit worse than the best place of heaven and closer one step to be deserving of being in hell[1].

In the following chapter, I will argue that just being too bad poses a problem to our MRP. I argue there that Callard's view is a good way to account for those problems. But I do not think that it shows any problem *in* Callard's view. As I said before, I suspect that most people who blame Hitler and praise Gandhi may find the hell-heaven view the first choice for representing their conceptions of blame and praise. But resenting Hitler can be put consistently in the Callardian view too. Suppose someone that because of some skeptical arguments is wholeheartedly against the hell-heaven view. This person still may think that she has good reasons to be (very) angry with Hitler just because of how he has diminished the moral community. Even if this person *also* thinks that she has reasons to *express* her (or the whole moral community's) anger to Hitler, she may insist that we should find other ways for expressing our anger to him besides making conditions for Hitler worse—for example, if we could imprison Hitler, his conditions in the jail should *not* be worse than other prisoners.

Conclusion

Our MRP and our ideas of when and how one should be held responsible are prima facie strong and robust. I tried to see *how* strong and robust it is. For this, I studied Strawson. I

[1] Galen Strawson's argument on the impossibility of (ultimate) moral responsibility attacks assuming a sense of responsibility for one's self that some philosophers like Fischer (2006) think is too inflated and call it, "metaphysical megalomania". One may think that the heaven-hell view of desert is only at issue when such a strong assumption is under scrutiny. I disagree. In chapter four, I suggest an argument structurally similar to Strawson's that I hope is relied on a moderate sense of responsibility for one's self that if works denies a sense of desert close to the heaven-hell view. Here, like what I said about P. F. Strawson, I only try to suggest that the heaven-hell view is part of the "picture" of MRP at least for many non-philosophers.

think a weaker Strawsonian claim is compelling. That is to claim that before understanding our MRP and the internal reasons governing it thoroughly, we should not be worried about the external reasons for or against the justification of these practices—like the reasons from the truth of determinism. However, for a stronger claim according to which the intuitive exempted/responsible distinction is strong enough to mark such external reasons as not relevant to the justification of these practices, both an opponent and a proponent has not an easy job to make a successful move. The debate is ongoing research. I did not discuss it. What I did was to suggest one necessary question, namely, how strong the sense of desert that Strawsonian reactive attitudes have to presuppose is.

CHAPTER 2:

THE PROBLEM OF BLAMING A SUPERBAD PERSON

Introduction

In chapter one, I argued for a weaker Strawsonian idea that our moral responsibility practices (MRP) and our ideas of when and how one should be held responsible are strong enough to support the priority of understanding these practices and the internal reasons governing them thoroughly to the external reasons for or against the justification of these practices—like the reasons from the truth of determinism. In this chapter, I try to show a problem that arises from two seemingly innocent assumptions in our MRP, namely the exempted/responsible dichotomy and what I introduce as the Direct Correlation Assumption of Angry Blame. The latter is the idea that more severe wrongdoings, when one is not exempted, make the wrongdoer deserving of stronger anger. This problem, I shall conclude, is a result of an overfocus on morally mediocre people and the moments that we/they think they can do the right thing only if they decide willfully enough. This overfocus leads to the problem of blaming a too bad person that I explain, mainly, in the case of narcissism.

I. The Direct Correlation Assumption of Angry Blame

In chapter one, I tried to show the centrality of understanding blame as a kind of anger that the blamer holds the blamed person deserved. Because of the centrality of this picture, those theories of blame that argue against either including anger in their accounts of blame

or the desert base of this anger will have a further task of explaining either how their account is still not revisionary or why we should replace our deserved-anger notion of blame with their revisionary suggestion. Pereboom (2013) and Scanlon (2008) are salient examples of such theories.

There is another assumption that comes as a following to the assumed centrality of deserved-anger in a common-sense notion of blame and might seem even more innocent than it: the assumption that if your being wronged by me gives you reasons for anger with me, stronger wrongs make the reasons of anger stronger. I call it the Direct Correlation Assumption. I try to reveal the Direct Correlation Assumption in Susan Wolf's criticism of Scanlon's account of blame. Briefly, Scanlon suggests that blame is the adjustment of our relationship with the blamed and the revision of our attitudes toward him in response to attitudes expressed in his wrong behavior (Scanlon, 2008, 122-3; 233n54). Susan Wolf criticizes Scanlon's account by pointing out that she gets angry towards her daughters when they use her shoes without telling her. She thinks this anger is reasonable at the same time that her daughters do not show any attitude that impairs their relationship with her (Wolf, 2011). However, I think Wolf's daughters *do* show some attitude that impairs their relationship with their mothers—attitudes like not respecting their mother's privacy, or her right to be sure that her shoes are there whenever she needs them, etc. Wolf's relationship with their daughters seems to her unimpaired because the extent of the impairment is not *noticeable*. If her daughters do some more similarly mundane things, the relationship can get worse noticeably. Wolf may insist that this example still is a good case against Scanlon's accounts because it is an example of some noticeable anger that reasonably targets some unnoticeable impairment of a relationship. Based on this (re-)formulation of Wolf's criticism, we can see that it relies on the Direct Correlation Assumption.

In this chapter, I try to show that the Direct Correlation Assumption is not as innocent as it might seem. I will discuss four reasons respectively: First, being too bad makes blame unintelligible for the blamed. Second, being too bad makes sadness appropriate. Third, blaming superbad people angrily has morally undesirable outcomes. Fourth, and connected to the third reason, if we suppose we have reasons for being angry at superbad people, we may have also to suppose that we have reasons for being angry with them forever. Some personality disorders like narcissism or some political and social problems like deep-seated

racism can be the cases of this problem. They are not exempted (even if psychopaths are, as I think they are), but blaming them, even without expressing it to them, seems not a (morally/rationally) right thing to do.

II. Narcissism as being superbad

In this section, I try to suggest just being superbad as a problem for our MRP. Some Personality Disorders (PDs) are instances of being superbad. The Diagnostic and Statistical Manual of Mental Disorders (DSM) defines most mental disorders based on some cognitive impairments and/or some deficits in emotional or behavioral self-control. Unsurprisingly, most philosophical studies about the significance of such mental illnesses for our MRP are focused to see whether such impairments/deficits are exempting/excusing (Kozuch and McKenna, 2015; Sripada 2015). But some PDs "are explicitly defined and diagnosed in part via traits that count as failures of morality or virtue (Pickard 2011).

The DSM's diagnostic criteria for some personality disorders are some moral traits that we might think that if we can reasonably attribute them to a person, and if we know that that person has no exempting (or excusing) condition—like having those traits as the result of a serious manipulation or some neurobiological abnormalities he has born with—, then, there would be no doubt in blaming them for those actions of them that express such personality traits. I call these personality disorders, Moral PDs.

The problem of being superbad is distinct from both the problem of psychopaths and the problem of being superbad because of some horrible upbringings. Although Hare's Psychopathy Checklist-Revised consists of morally significant traits, another part of the dominant picture of psychopathy—in both the psychiatry and philosophy domains—holds psychopathy as being based on some substantial hardware—say, brain—problems[2]. Unlike

[2] The views on the moral responsibility of psychopaths can be grouped into three views: First, Watson (2011) who thinks responsibility has two "faces" of attributability and accountability and holds psychopaths responsible in the first and not responsible in the latter. Second, those like Levy (2007; 2011: 208), Nelkin (2015), and Jaworska (2017) who deny the responsibility of psychopaths in both senses—*if* there are two such senses. Third, Scanlonians like Scanlon (2015), and Angela Smith (2015) who think there is only one (blame- or praiseworthiness) responsibility and on which psychopaths are responsible. The hardware view of psychopathy can be seen as a vital assumption

psychopathy, the problem of horrible upbringings can be a software problem. The most discussed case of this problem is Gary Watson's Robert Harris (Watson, 2016). Harris is a cool murderer and *is* superbad, but his case is puzzling because of our contradictory reaction to his being now superbad and his harrowing and abusive upbringing. But I am trying to show the problems that just being superbad may show in our natural ideas of MRP. To use Watson's words, my focus is on the "inexplicable" case of "bad apples" who are "just *as* vicious" as Harris but have "had a supportive and loving environment as a child (ibid)."

The most notable Moral PD is narcissism or NPD that all its diagnostic criteria are things overtly moral like a grandiose sense of self-worth, being too envious, unwillingness to sympathize with the needs and pains of others, being not open to any criticism, etc.

Narcissistic Personality Disorder

Diagnostic Criteria	**301.81 (F60.81)**

A pervasive pattern of grandiosity (in fantasy or behavior), need for admiration, and lack of empathy, beginning by early adulthood and present in a variety of contexts, as indicated by five (or more) of the following:

1. Has a grandiose sense of self-importance (e.g., exaggerates achievements and talents, expects to be recognized as superior without commensurate achievements).
2. Is preoccupied with fantasies of unlimited success, power, brilliance, beauty, or ideal love.
3. Believes that he or she is "special" and unique and can only be understood by, or should associate with, other special or high-status people (or institutions).
4. Requires excessive admiration.
5. Has a sense of entitlement (i.e., unreasonable expectations of especially favorable treatment or automatic compliance with his or her expectations).
6. Is interpersonally exploitative (i.e., takes advantage of others to achieve his or her own ends).
7. Lacks empathy: is unwilling to recognize or identify with the feelings and needs of others.
8. Is often envious of others or believes that others are envious of him or her.
9. Shows arrogant, haughty behaviors or attitudes.

Table 1 - NPD's Diagnostic Criteria in the DSM-5

in the arguments of Watson, Levy, and Nelkin, and, at least, is not denied by Scanlon and Smith. Although Levy's and Nelkin's arguments seem convincing to me, I do not discuss this debate.

It is important to note that the narcissist's bad personality traits are not *caused* by his mental disorder, they *are* his mental disorder. In Nomy Arpaly's words, narcissism is not "just like diabetes":

> we cannot truly say ''it's not that he really is indifferent to the needs of others, it's only a disease like diabetes,'' or ''it's not that he really thinks he is superior to you, it's only a disease like diabetes.'' The narcissist *does* think he is superior to you. He *is* indifferent to the needs of others. […] As a result, labeling someone with the official term ''Narcissistic Personality Disorder'' is hardly any less derogatory than simply calling someone ''a narcissist'' or ''a selfish, self-absorbed megalomaniac (2005: 289-290).''

Although I think Arpaly's view here is both correct and important, I think any mental disorder *is* like diabetes in another aspect, namely, both cannot be wished away: trying to 'snap out' of any mental disorder is as futile as trying to 'snap out' of diabetes. This is one of the points, Arpaly starts and ends her paper with. I discuss Arpaly's account of moral responsibility in more detail in the next chapter. However, as far as I know, Arpaly has never discussed how, if any, the fact that narcissism or self-absorbed megalomania cannot be snapped out of can be important from a responsibility/blame point of view. The only philosopher who has discussed Moral PDs from a responsibility/blame point of view is Hannah Pickard. I study her in the next section.

III. Pickard's theory of "responsibility without blame"

Hannah Pickard sets up the blame/responsibility problem of Moral PDs based on three assumptions: 1) PD is not an exempting condition, and on many occasions is not also an excusing condition. 2) Some important parts of PDs *are* moral. As a result of 1 and 2, PD can be something that can be blameworthy or can make the person do many blameworthy actions. 3) But blaming people with PDs, by clinicians, can impair the effective treatment by decreasing the quality of care they can receive and by making them discontinue the process of treatment. This makes a "clinical conundrum" that Hanna Pickard (2011, 2013) has tried to develop a theory for resolving. However, I think Pickard's theory is at best incomplete and at worse inconsistent. She criticizes "detached" accounts of blame, like Scanlon's, for not covering, what I introduced before as, the deserved-anger. Surprisingly,

the main claim of her theory is that we should replace our deserved-anger included notion of blame—that she calls, "affective blame"—with "detached blame" only for people with PDs while she argues in detail that we have no reason for any significantly different judgments about their responsibility.

Pickard's theory has two parts: a "conceptual" and a "practical". In the conceptual part, she argues for five claims: 1) People with PD are responsible and should be held responsible—they are not excused in many cases, 2) Blame comes in two forms: detached vs. affective, 3) Detached accounts of blame are substantially incomplete, 4) But only by blaming people with PDs in a detached way the conundrum can be solved, 5) So, we should hold people with PDs responsible, blame them in a detached way, and avoid blaming them in an affective way. Pickard thinks a practical part is needed to show how clinicians can, in practice, avoid affective blame as affective blame is our natural way of blaming people who we held responsible. I think Pickard does not resolve the problem. What she, in fact, does is to wear a new cloth to the wrongheaded idea that just because some moral traits are labeled as "mental disorders", they deserve different reactions. In what follows, I explain these five points.

Pickard is concerned not just with NPD. Her concern is any PD that is defined, at least in part, by the morally significant personality traits. To point this out, Pickard refers to DSM-IV-TR (1994) definitions for some PDs. Her instances are:

> Narcissistic PD as involving lack of empathy, grandiosity, need for admiration, and a willingness to exploit others. Histrionic PD involves an excessive demand for attention and 'inappropriate' sexual behavior. Borderline PD involves extreme and inappropriate anger toward self and others, instability in self-image and interpersonal relationships, and marked recklessness, impulsivity, and paranoia. Antisocial PD involves disregard for others and violation of their rights, criminal behavior, and lack of remorse […] Paranoid PD involves unjustified suspicion and distrust, and a tendency to hold grudges against others. Obsessive-Compulsive PD involves forsaking friendship for productivity, obedience to rules and authority at the expense of the good of self and others, miserliness, stubbornness, and a desire for interpersonal control (2011).

Pickard sets up the problem with these PDs by showing that an effective treatment for a person with a PD "both presupposes and fosters" her capacities as a responsible agent, and at the same time blame is detrimental to such treatment.

She begins by explaining why people with a PD should be held responsible. Her first reason appeals to what she calls our "common-sense conception of agency (212)". As she says, people with PDs "on at least most, if not indeed all, occasions," are informed about the moral significance of their actions and can control their behaviors to do otherwise according to this knowledge if they are motivated for it. Her second reason for holding people with PDs responsible is practical. She points out that if clinicians give up the belief that people with PDs are responsible, all psychological treatments lose their point, "leaving only medication as an option (214)". It is because a psychological treatment for a PD is "in part, augmenting service users' existing capacity for agency (213)".

After showing that why people with PDs *are* morally responsible agents, and why clinical practices *should hold* them responsible for their behaviors, Pickard tries to show why blame is detrimental to the effective treatment of PDs. She thinks blaming attitudes in clinicians can be responsible for the less amount of care they receive in mental health centers—according to the reports she cites. Blaming people with PDs also increases the risk of disengaging with the treatment in them. This is why she suggests, a "responsibility without blame" theory.

So far, Pickard has told us about the responsibility and the blameworthiness of people with PDs. In the last part of her conceptual framework for the possibility of responsibility without blame, she tries to distinguish two kinds of blame, "affective" and "detached":

> Detached blame consists in judgments of blameworthiness, and may further involve correspondingly appropriate revisions of intentions, the imposition of negative consequences, and accountability and answerability. […] Affective blame consists in negative reactions and emotions, whether rational or not, that the blamer feels entitled to have.

She agrees that detached accounts can cover some important parts of blame, but she thinks such accounts cannot explain two facts about blame: the possibility of irrational blame, and the "sting" that the blamed person generally feels. In other words, Pickard believes that a

plausible account for the nature of blame should be based on the affective blame, not on the detached blame. But she also thinks that detached blame "can have a place within effective clinical treatment, and, insofar as [it] encourage[s] responsible agency, may be essential to it." But "affective" blame is harmful to the treatment of PDs and hence we should avoid it for people with a PD. So, responsibility without blame is in fact "responsibility without affective blame, […] no matter what the service user has done (Pickard, 2011)."

Because of holding the affective and not the detached blame as representative of natural blame, when she finishes her job on the conceptual framework of her theory, Pickard adds a "practical" part in which she suggests how clinicians in practice can avoid the emotional parts of affective blame to people with PDs and restrict blaming them to detached blame. She thinks clinicians must "cultivate" compassion and empathy because these emotions "simply cannot comfortably coexist" with the negative emotions of affective blame. For cultivating empathy and compassion, she suggests paying "proper attention" to the history of people with a PD, because she thinks these people "often come from harrowing backgrounds, impoverished of all goods." She mentions some of such harrowing backgrounds:

> dysfunctional families, where there is breakdown, death, institutional care, and parental psychopathology; traumatic childhood experiences, with high levels of sexual, emotional, and physical abuse or neglect; and social stressors, such as war, poverty, and migration.

Having explained Pickard's theory, now I can argue for my claim that Pickard does nothing more than wearing a new cover to the wrongheaded idea that just labeling some moral traits as "mental disorders" makes them deserving of different reactive attitudes. First, Pickard's list of harrowing backgrounds is not part of DSM's criteria for Moral PDs—most notably, NPD. Second, even for those people with PDs with such harrowing backgrounds, Pickard's reason for why they should not be blamed by the affective blame is not their harrowing backgrounds. The harrowing backgrounds for Pickard—unlike Gary Watson's Robert Harris case that I study later—are just a practical tool. Pickard insists on this point by asking us

> to recognize that this appeal to past history does not eliminate responsibility or blameworthiness. It may reduce responsibility, insofar as certain kinds of background impede the development of skills that, for instance, facilitate emotional regulation and, correspondingly, behavioral control. Equally, extreme impoverishment can limit choices, which can sometimes excuse bad decisions and the harm they cause. But such reduction is not global, and depends on the particular kind of background, skills, choices, and harm in question (2011).

In other words, Pickard's "responsibility without blame" theory can be summarized in two theoretical "although" and one practical goal that eventually lead to a "should": 1) although PDs in themselves are not exempting/excusing conditions, 2) although some behaviors that are constitutive of the personality of a person with a PD *are* blameworthy, but 3) [just] because the effective treatment of PDs, as a valuable practical goal, needs avoiding some essential parts of natural blame, we should avoid those parts. I think this inconsistency in Pickard's theory is a result of some substantial incompleteness that cannot be filled without taking the claim of this chapter seriously. I try to argue that just being too bad is problematic for our natural ideas of MRP. I am sympathetic with Pickard's view to the clinical consequences of blaming Moral PDs. But I think this is true for *all* cases of being superbad. Consider a society with some extreme levels of racism. The racist people may have no excuse and then should be held responsible. But blaming them angrily makes things just worse. One might think that it only shows that we should blame them privately. But I think what I quoted from Watson at the end of the last chapter suggests that there is something more than practical caution in Gandhi's and Martin Luther King's holding their opponents responsible with some kind of love instead of the deserved-anger. As I understand (Watson's reading of) them, they think blaming their racist opponents angrily should be abandoned because it is based on our ignorance of some deep facts about human nature that have made them the kind of racists they have become.

I think the claim of this chapter is the view behind a general idea in DSM, according to which any other person with "an enduring pattern of inner experience and behavior that deviates markedly from the expectations of the individual's culture, is pervasive and inflexible, has an onset in adolescence or early adulthood, is stable over time, and leads to distress or impairment pervasive, inflexible, and maladaptive personality patterns" should

receive the psychiatric cares people diagnosed with a PD receive even if "the criteria for any specific personality disorder are not met (DSM-5: 645)".

IV. Arguments for the problematic nature of blaming superbad people

I think blaming the narcissist and the racist opponents of Gandhi and King is a substantive problem that is overlooked as a consequence of two things: 1) an overfocus on morally mediocre people and the moments that we/they think they can do the right thing only if they decide willfully enough—call it for short, mediocre blameworthiness. 2) what I introduced as the Direct Correlation Assumption. But maybe mediocre blameworthiness is the exception, not the rule. In this section, I start by suggesting one way that may explain why angry blame in cases other than mediocre blameworthiness may lead to some substantial problem. My suggestion draws on Agnes Callard's argument for when we have reason to be not angry forever. Then, I argue for a direct argument for sadness for the claim of this chapter.

In "The Reason to Be Angry Forever", Agnes Callard (2017) tries to solve a problem from "the eternal anger argument". According to the eternal anger argument, if your reason to be angry with me is the fact that I have wronged you, as the facts in the past remain valid forever, you have reason to be angry with me forever. She, first, rejects a common solution. The common solution is based on a common understanding of why a person who we have reason to be angry towards—call this reason, R1—can do something that can give us some good reason like R2 to cease our anger towards her. The common understanding that, as Callard shows, goes back to at least Aristotle, holds R1 as a desire for bringing the wrongdoer to give us R2 by things like an apology. Callard argues that this picture does not match with the phenomenology of anger like when anger makes us not interested in our previous relationship with the wrongdoer anymore. The main error Callard diagnoses behind this picture is that "it treats the angry person as too autonomous, self-possessed, and too aloof from the damage the other has inflicted on him or her (135)". Against this picture, Callard argues that you have R1 reason to be angry at me because 1) before my wrong, we had been valuing something together, and 2) "I refrain from holding up my end

of the valuational burden, making it impossible for you to hold up yours in any way other than anger (131)." It will be easy to guess that Callard holds R2 as whenever I do something like an apology that can give you reason that we can back to our co-valuing relationship again.

If Callard's argument on the correct nature of reasons for anger is compelling—as I think it is—, it shows a conditional claim, namely, if you suppose you have reasons for being angry with a superbad person, you will probably have those reasons forever. It is because, Callard's argument shows, as long as a person does not come to understand her role in worsening her relationship with others, there will be no R2 for ceasing our reasons for anger with him. It, then, explains why being angry with Gandhi's and King's opponents or narcissists may be problematic.

One might say that there is no problem in having reasons for being angry with a person like Hitler forever. I agree partially and will account for this intuition in my positive account that I will defend in the last chapter. I agree that *one part* of the reasons for blaming a wrongdoer is to declare his valuings as bad. But there are other parts that Hitler's example is not good for seeing them. It is because he and his proponents are dead now, a fact that is not true about Moral PDs and the deep-seated racism that Gandhi and King were concerned about. Eternal anger with these non-dead superbad cases will be a practical but substantial problem because angry blame may mean having less Gandhi and King and more Che Guevara and being less willing to look for the roots of making people superbad and for the effective social and psychological treatment.

Another explanation for the problem of blaming superbad people that I have no time to discuss in detail is the problem of intelligibility. According to one necessary condition of Watsonian accountability, at least one base for the reasonability of blame is that it should be intelligible for the blamed person. But as Watson says, "[i]n what sense can a deeply cruel man respond to reasons of kindness?" (Watson, 2016)[3]. I think this is a powerful intuition that is asymmetrically true also about good people. A person who does not feel any force in the reasons of bribery—like "if you get the bribe, you will have a much easier life and the risk of being caught is very low"—may be a better person than someone who

[3] This sentence does not occur in the 1987 version of Watson's paper.

has to think a lot for coming to refuse the bribe. I hope the account I suggest in my final chapter explains this intuition.

Finally, I think a direct argument from (moral) sadness can support the claim of this chapter. Imagine a teacher who cares about establishing good relations with her students. But she finds that some of her students abuse her trust. Suppose she tries her best to show the value of strong relationships to these students but always fails. Suppose also that she finds that these students have no deep friendships and all their friendships are for short-term joys and benefits. I think it is not surprising if this teacher feels sad about these students even if she does not find any hardware disorder like psychopathy or some dramatic childhood true about them. Although we do not expect all people to be as caring as this teacher, we may admire her and not criticize her as too sentimental. As accommodating for this sadness in our natural ideas of MRP is, at least, a philosophical task, I think this direct argument from sadness is stronger than its length may suggest for the claim of this chapter.

The problem may be solved by denying the Direct Correlation Assumption. But denying this assumption is not so easy. Recall Susan wolf's anger at her daughters. Certainly, she has no reason for anger to very small behaviors of her daughters like if they move her shoes just a bit to free space for their shoes. On the other hand, she may have reasons for some stronger anger when her daughters use her shoes without her permission again after seeing how this makes their mother angry. So, the strength of anger one may have reasons for *is* directly correlated with the strength of the wrongdoer's wrongdoing to some point. But there is a point from which this direct correlation starts to seem implausible. After a few times of blaming her daughters angrily without the result she wants to see, Susan Wolf has reasons to let her anger aside and look for some deeper problems in her daughters or her relationship with them. I try to account for this intuition in my positive account in chapter four.

One may object that there are some actions like child abuse and torture that our intuitions are very different about them. We may expect it less straightforward for an apology to cease our reasons for anger at their agents, we do not depend our blame on the intelligibility of the reasons contrary to those these agents have acted upon, we do not expect a caring person to feel sad about them, and we feel the Direct Correlation Assumption to be robustly

correct in their case—that is to say, more repeated or harsher child abuses or tortures warrant stronger angry blame. I think the different intuitions here are because of the disgusting nature of these bad actions. This disgusting nature of these actions is not, I think, because they are necessarily worse than many other wrongdoings. Fraud elections or many other immoral ways of gaining political power can be as bad as or even worse than torturing the opponents, say, when they predictably lead to dramatically decreased quality of lives and increased deaths. The same may be true for educating a child to be a terrorist or a mafia leader[4].

Conclusion

I have argued that being superbad when it is not because of any exempting condition is problematic for the deserved-anger view of responsibility that is, as I argued in chapter one, is the natural view behind our MRP. I argued that angry blame loses its intelligibility for a superbad person. If we suppose we have reasons for anger, we have to suppose that we have reasons for being angry at a superbad person forever. I suggested that it has serious morally undesirable outcomes. I also argued that just being superbad makes sadness appropriate. I think that these problems all are the results of some substantial deficits in our understanding of morally responsible agency that is a result of an overfocus on morally mediocre people. In the final chapter, I suggest a positive account of responsibility that I hope will remedy these problems.

[4] I defended a negative claim on disgust, namely, against a necessary correlation between being (morally) disgusting and being morally bad. There is a quite wast literature on the positive views on disgust. For a review of this literature see: Strohminger, 2014.

CHAPTER 3:

ONE'S QUALITY OF VALUING

Introduction

In chapter 1, I described the general idea I follow in this thesis. The idea, following Strawson, is to prioritize the problems *internal* to our MRP. In chapter two, I suggested that our MRP face a problem when we need to know how to react to people who are just too bad. The question of what kind of reaction—sadness or resentment (or both)—can be justified as concerns this level of non-exempted wrongdoers depends on, I suggest, the question of what *kind* of responsible agents we can and we can't be. In this and next chapters, I try to answer this question by describing us as (co-)valuing creatures.

As (co-)valuing creatures, I argue, we develop our valuings through, what I call the Responsibility Chain of our lives: We are responsible for those actions and choices, or lack thereof, when they express our valuings. But our valuings can be *our* valuings only when they're partly the results of our previous actions and choices. So the constitution of our valuings is, directly or indirectly, the result of good and bad luck. I argue that this necessitates understanding blame (or praise) as compatible with and connected to moral sadness (or counting the praised as lucky).

My argument for this view draws on important expressivist ideas about moral responsibility put forward by T. M. Scanlon, Nomy Arpaly, and Angela Smith. In this chapter, I defend their view that responsibility judgments can go beyond one's conscious

values. I argue that this can only be true if the unconscious values have not always been sub-personal. In the next chapter, I criticize the history-insensitivity of expressivists and suggest a way for making them history-sensitive.

I. The expressivists' view

One of the most important families of views on what responsibility judgments are about is suggested by Scanlon (2008), Arpaly (2002), and Smith (2005). Their view is sometimes called the quality-of-will-based theory (Arpaly, 2002: 115) and sometimes called expressivism (Levy, 2011). I see their core idea as the Expressivist Thesis:

> When we judge someone's moral responsibility, we judge whether an action, desire, pattern of thought or feeling, or any other mental state, or lack thereof, is expressive of attitudes that are ideally or normatively sensitive to her evaluative judgments (or, for Arpaly, of the depth of her moral concern).

Their view is a descendant of P. F. Strawson's quality of will view and updates the influential views of prominent moral-responsibility philosophers like Harry Frankfurt and Gary Watson. Expressivists provide a clear starting point for what we should look for in our judgment of someone's responsibility; that is why, when, and how things like one's knowledge and control should or shouldn't be seen as important in our judgment.

Consider coercion. Scanlon gives an example of a bank teller who gives cash to a robber to avoid a deadly threat. As Scanlon explains, coercion here shouldn't be seen as exempting the teller from moral responsibility responses. Rather, it shows different reasons the teller has acted on, reasons that might make her the target of a different blame, or even of praise if it shows that she's managed the situation wisely (Scanlon, 2008: 180-1). Or consider awareness. Your best friend might be entitled to blame you if you forget her birthday. Your unawareness of the date can't be an excuse. The blame here is for your very unawareness of the date because it is a result of not valuing your friendship enough. What can be an excuse, then, is a different explanation, like how exceptionally busy you might have been around the date (Smith, 2005: 236; 248, fn. 21).

Expressivists also argue for, what I call, the Beyond-Consciousness Thesis and the History-Insensitive Thesis. Although for them, these theses may be not separatable from the Expressivist Thesis, I think they are distinct theses. According to the Beyond-Consciousness Thesis, responsibility judgments can go beyond one's conscious values. According to the History-Insensitive Thesis, the way by which someone has come to her conscious and unconscious values doesn't change her responsibility for having those values. I will now explain these two theses in more detail.

Scanlon (2008: 195) and Smith (2005: 260-1) discuss the example of a sincere anti-racist to whom some negative thoughts about people of certain races occur frequently. As Scanlon explains, her sincere, explicit anti-racism and her shame for the unwanted thoughts affect our overall judgment of this person but she is still responsible for the unwanted thoughts. Precisely because of this responsibility, she is, and should be, ashamed. Arpaly's most famous example is about the praiseworthiness of Huckleberry Finn for not being able to turn in Jim, his runaway-slave friend, even when it goes against his best, conscious or "official", judgment that helping an escaping slave is stealing and wrong (2002: 76-79).

To understand expressivists' History-Insensitive Thesis, we need to understand what they think is involved in holding someone morally responsible for an action or attitude. The main thing involved in holding someone morally responsible is not the same for Smith, Scanlon, and Arpaly. For Smith, it means "that it would be intelligible to ask her to 'answer for' that thing—to give her (justificatory) reasons for thinking, feeling, or acting in the way she has […] and this is the key to opening the door to the further moral responses" (Smith, 2015: 103). This is what Smith calls Answerability and argues to be "the most basic sense" of responsibility (ibid.). For Scanlon, the prerequisite of thinking about the philosophical controversies about freedom and responsibility is to know what blame is. Scanlon suggests that blame is the adjustment of our relationship with the blamed and the revision of our attitudes toward him in response to attitudes expressed in his behavior (Scanlon, 2008, 122-3; 233n54). This is why Scanlon calls his account of the sense of responsibility that is the basis for blame- and praiseworthiness, "moral reaction responsibility" (2015: 89). Hence, I call his account of blame the "relationship-adjustment view." For Arpaly, blame is "a belief-like attitude similar to fear or various kinds of esteem [and] is analogous to holding someone to be a bad businessman or a lousy artist" (Arpaly, 2002: 172-3).

Now we can understand why expressivists are history-insensitive. Holding someone to be a bad businessman or a lousy artist is in itself insensitive to how she has become so. Or, for a relationship-adjustment conception of blame like not trusting an untrustworthy person, it doesn't matter "if he cannot help being untrustworthy" (Scanlon, 2008: 188). Similarly, looking for the evaluative judgments behind one's (non-)actions seems independently intelligible of how she has come to have those evaluative judgments.

I think expressivists' Beyond-Consciousness Thesis extends the domain of responsibility judgments plausibly, although I show in Section III that this thesis needs modification. The Beyond-Consciousness Thesis explains one important point of psychotherapy by which we are looking for how we really, even unconsciously, value things, persons, and ourselves. In section III, I discuss other cases in which we can come to *discover* about ourselves how we value things on which we never reflect.

Unlike the Beyond-Consciousness Thesis, expressivists' history-insensitivity denies some important intuitions. Consider how Arpaly discusses an example of brainwashing from Alfred Mele (Mele, 1995: 145). The example is about Beth, a philosophy professor who values philosophy much less than her colleague Ann, who works twelve hours a day, seven days a week. To make Beth work like Ann, some brainwashers instill the same hierarchy of values as Ann's in Beth and eradicate Beth's competing values. As Arpaly admits, there's a powerful intuition about a significant difference between the blame and praise of which Beth and Ann can be worthy of that "often serves as an intuition against which to check theories" (128). Arpaly's defense of the History-Insensitive Thesis is based on the assumption that Beth's change of values is similar to a variety of other changes we are okay with. Her examples include:

> shifts from being self-endorsed party animals to being self-endorsed industrious workers because of mysterious factors [one] regard[s] as "age" or the "drying up of hormones" […] begin[ning] to value parenthood—value, not just like—the moment their (formerly unwanted) children are born. [Replacing] atheism with "religion" (or vice versa) as the result of an experience of extreme loneliness and pain. (128)

Arpaly claims that "the only thing that distinguishes the[se cases] from Beth is that their irrational conversions *are not the result of a deliberate and wrongful action by another*

human being" (128; emph. in original). But Arpaly is wrong. There are cases that have no roles for other human beings where we react as we do to Beth. Consider someone whose change of values occurs magically because of a car accident, or during one night's sleep, and so on. The change here may be (morally) (un)desirable, for example, if a car accident makes someone a kinder person. Still, it doesn't change the blame or praise of which one is (or has been) worthy. Arpaly's cases, in contrast to these ones, the change might reveal that one has been worthy of different blame or praise than what we had thought. For, in her cases, the change can show the quality of one's valuings before the change. For instance, the atheism of someone who becomes religious only because of loneliness or pain (and nothing epistemically worthy), we can suppose has had a different epistemic worth than one who remains atheist in similar or harsher conditions, and vice-versa. But Beth's change of values doesn't tell us anything about how *strongly* she has been valuing a work-life balance. We know by my valuing reading of the Expressivist Thesis that an action or attitude is *up to an agent* when it is expressive of the agent's valuings. What I try to add here is that we need also to know when one's valuings are (or have been) up to her.

In the remainder of this thesis, I show how modifying the expressivist view, mostly by making them history-sensitive, provides an appealing answer to my question of what kind of responsible agents we are and of what kind of moral reactions this makes us an appropriate target. In the next section, I suggest a stronger reading of the Expressivist Thesis that changes its focus to how well or badly one (dis)values good and bad things. Then, I put aside moral responsibility in section III to show how ascribing valuings to someone can go beyond her conscious propositional attitudes. In next chapter, I ground my account of moral responsibility on my answer to this question.

II. One's quality of valuing as a better reading of the Expressivist Thesis

In this section, I suggest that the Expressivist Thesis is better to be read as a thesis about the necessary and sufficient conditions of ascribing someone good or bad valuing, that is, merely to claim that one (dis)values something (or doesn't (dis)value something) *and* to evaluate it as (objectively) good, bad, or neutral. I use "disvalue" as a stronger term than

"not value." For example, if you disvalue being my colleague, you see it as a bad thing, you have negative emotions toward it, or take it as a reason against working where I do. My understanding of valuing relies on Samuel Scheffler's account (2011), which is, independent of any responsibility debates, focused solely on when one values something. According to Scheffler, one values X when she is at least emotionally and reason-responsively disposed to a variety of X-related considerations and believes that X is valuable.[5] Later, I argue that the belief component can go beyond consciously held propositional attitudes. I show how this reading is a philosophically important modification to Scanlon's and Smith's views and more clearly expresses Arpaly's.

For Scanlon and Smith, evaluative judgments should be replaced with valuing. How a person (dis)values important things is more important than her beliefs about the valuability of those things for the reasons we have for adjusting our relationship with her or for (rationally) expecting what she effortlessly remembers or notices. As Scheffler shows, we can judge, sincerely, many things like Bulgarian history or opera-going as valuable without valuing them when they shape no noticeable part of our lives. Indeed, it's a basic fact about human beings that we can value a "tiny fraction" of what we judge as valuable mostly because we can't be emotionally and reason-responsively disposed to too many things at any given period of our life (Scheffler, 2011: 27). Valuing something differently when the judgment about its valuability isn't different can also occur in personal relationships. A friend may value her friendship with you more than another friend of yours simply because the former has spent more time with you and it has made her be more strongly disposed, emotionally and reason-responsively, to you and to things related to you. This difference can occur even when these two persons judge you as a valuable friend equally. Most reasons you have for adjusting your relationships with these two friends come from their different valuings rather than their similar judgments. Similarly, we don't, rationally, expect these two friends to remember and notice things important for us to the same extent.

[5] I omitted the further component of "seeing one's emotional susceptibilities merited" from Scheffler's account because I think the belief component implies, or even means, that one sees both one's emotional susceptibilities and, I add, her reason-responsiveness dispositions toward the valued thing merited.

My valuing reading gives clearer content to Arpaly's view. I show this in her explanation of Huckleberry's "perceptual shift" to "see Jim as a person" (77). As Arpaly explains, "Huckleberry constantly perceives data (never deliberated upon) that amount to the message that Jim is a person, just like him […]: equally ignorant, share the same language and superstitions" (77). But why does this unconscious perception occur to Huck and not to most people in his community? The most explicit engagement with this question is Arpaly's footnote with other examples of "perceiving fairly sophisticated truths without perceiving that one is perceiving them. [Like] when the confession of a cheating spouse is surprisingly unsurprising" (ibid.). But again, we can ask *why* these unconscious perceptions happen for some and not for others.

In my argument for replacing evaluative judgment with valuing in Smith, I showed the possible role of one's valuings in what she notices or neglects. Similarly, unconscious perceptions are likely to happen, and matters in our judgment of a person, only when the object of the perception is something the perceiver (dis)values, say, her marriage or friendship.

III. The personal-level condition of valuing

Recall that valuing is composed of a cognitive component and emotional and reason-responsiveness dispositions. In this section, I elaborate on the cognitive component. I study Levy's Consciousness Thesis (CT) to show that while it can't show that one's valuings should be based only on her present or previous consciously propositional attitudes, a present or previous cognitive relation between valuer and her valued object at the personal level is necessary for valuing. I argue that there are cases in which the personal-level condition can be met while the consciousness condition cannot.

Levy denies the blame- or praiseworthiness of cases like Smith's forgetting of a birthday and Arpaly's Huckleberry. His argument depends on what he calls the Consciousness Thesis: It is a necessary condition for being responsible for something that one is or has been consciously aware of its moral significance. Levy's defense of the Consciousness Thesis against expressivists such as Smith and Arpaly can be put in three premises: 1) only conscious facts are globally broadcast in a person's mind; 2) for a fact to show one's

evaluative judgments (or moral concern), it should be possible for one to evaluate the fact against all her relevant evaluative judgments; and 3) any fact that is not globally broadcast in a person's mind cannot be evaluated by her against all her relevant evaluative judgments (Levy, 2014: 90). We can see in this description of the Consciousness Thesis that it's a thesis for suggesting a necessary condition for ascribing valuing before and independent of being a thesis about moral responsibility. So in this section, I leave aside moral responsibility and read the Consciousness Thesis as being about ascribing valuing.

Let's start with Levy's negative claim in the birthday example. I think Levy's argument for the cases like the birthday example works only as an epistemic warning. The blame in these cases is based on a conditional claim like: You would remember—be conscious of—your friend's birthday if you valued the friendship more. The blamed person here is blamed for being *not* conscious of the relevant fact, the date of her friend's birthday. Negative valuing claims like this will be wrong only if the conditional is wrong. However, in Smith's birthday example, the conditional seems true: The importance of a trip, a visiting research period, and so on in one's eyes are important factors—among others like how busy her life is, how well her memory works generally, and so on—for determining how likely it is that she will remember the relevant details. So Levy's point can be plausible only as an epistemic warning to check other possible explanations before concluding a deficit in the blamed's valuings. I think Smith has anticipated this warning when she talks about the "normal" connection between one's remembering or noticing and one's evaluative judgments, that is, when there's "no [other] apparent reason" (Smith, 2005: 248n21).

But the more important point of disagreement between Levy and expressivists is on implicit attitudes. I think the appealing force of the Consciousness Thesis is in fact from a weaker thesis: Ascribing valuing shouldn't be based on those mental states that have been always sub-personal—call it the Sub-personal Thesis. I modify one of Levy's examples to make it a case for which the Sub-personal Thesis is most clearly true. He reports a test in which people are asked to select a police officer from candidates with different degrees of street wisdom and education. The results clearly disfavor females, though the reasons the participants provide for and sincerely think of as supporting their choice are good in themselves (2014: 93-4). The participants don't think they dislike seeing females as police officers. There is something *sub*-personal in them that does. To make this case a clearer

example of always-been sub-personal attitudes, suppose that the only reason behind the participants' bias against females is that sometime before the test, we have shown them subliminally scenes in which female police officers do some unpleasant things. The implicit attitudes here have no weight for any valuing claim about the participants, such as not valuing gender equality. The bias emerges solely because of their sub-personal nature. This result can be also true if the effect of subliminal manipulation lasts much longer. Suppose, for instance, that the subliminal pictures are shown to the participants every morning, via the monitors of the subway they take, and that their effect lasts the whole day. These people's valuings—like how they value gender equality—can remain intact as the manipulation may make them biased only to female police officers and not to female politicians, CEOs, etc.

A caveat may be needed here. I assume the lack of any opportunity for the subliminally manipulated cases above for coming to see their bias against female police officers. When there are such opportunities but the person doesn't use them, it may show how she values *other things* like self-reflection, reasonable objections of others, etc. In real life, the conditions I made here for when implicit attitudes can't show one's valuings just because of their sub-personal nature is very unlikely to happen. Still, as I try to show in the next chapter, these cases might show a new argument for the claim that expressivists' account of responsibility is too broad because they can't exclude the cases I introduced here.

Unlike the above example, there are cases of implicit but not sub-personal attitudes that I introduce as Eluding cases and Default cases. This is, I think, the main point overlooked in Levy's argument against expressivists. Levy's argument is supported both intuitively and empirically. The empirical shows that only conscious attitudes are "broadcast to the full suite of the consuming systems that drive action" (80-81), and then concludes that implicit attitudes are sub-personal because they are "available only to some modules" (83). But there are important cases of implicit attitudes that are intuitively personal. Intuitions are of special importance here because the personal/sub-personal distinction is understood by Levy, and in the literature on the distinction (see: Elton, 2000), by appeal to when we think intuitively that something can "rightly be predicated of the person herself" and when only "of some lower-level components of the mind [like] 'edge detection occurs in V1'" (Levy, 2014: 31n8). On the empirical side, Levy doesn't cite any empirical study for the kind of

cases I'll discuss and, I suggest, empirical studies for these cases may not be even possible, at least now.

In an earlier book, Levy's argument was against grounding (the valuing base of) one's moral responsibility on implicit attitudes only when they conflict with her explicit attitudes (Levy, 2011: 189). The Consciousness Thesis is more general. The important point dropped in Levy's later book is to account for the possibility of cases in which one *discovers* what she has been (dis)valuing, (or lack thereof) possibly for a long time. Many philosophers—including Smith (2005: 252), Levy (2011: 189), and Scheffler (2011: 38)—have talked about this possibility. Indeed, it is almost the whole point of self-reflection with or without the help of a friend or therapist and hence is worth keeping. In the remainder of this section, I argue that a personal-level cognitive relation between valuer and (dis)valued object can be met, at least, in two kinds of implicit cases: Eluding, and Default.

Consider a person who repeatedly enters into humiliating relationships or a person who always studies or works hard but in fields with nothing in common between them. It's quite possible that a claim about these persons made by a therapist or a friend may be true that they *see* themselves as unworthy, that the humiliating relationships are to prove this to themselves, and that working hard is just done to escape from this, implicitly claimed, 'fact'. In these Eluding cases, one *disvalues* oneself unconsciously. A similar claim is true about Huckleberry. He *values* his friendship with Jim and it makes him see Jim as a person. But the valuing and the resulting perception "*eludes* him," not only because he hasn't the "genius of John Stuart Mill" (Arpaly, 2002: 77; my emph.), but also because he has been valuing testing the moral/religious views of his community less than a person like Mill.

The second, Default cases, appear when one sees (always) something as (dis)valuable unconsciously because it is the default way one sees that (dis)valued thing without ever reflecting on it. I was at a talk by a graduate student, call her Janaan, advocating the claim that human and nonhuman animals are equal in moral standing. She predicted that there will be a time in the future when humans see her opponents as how we now see those who supported slavery. Imagine some audience members who sincerely think they don't see humans as superior because since they can remember they have loved pets and human children equally. However, some of these audience members might admit that they'd save

the life of a human rather than a pet if they have to choose only one. For many of them, this reflection doesn't *generate* the belief that humans are superior to nonhumans; it only makes it explicit. Still, we can say that these people were not valuing the equality of human and nonhuman animals even before this awareness.

To conclude, I think in our intuitions Eluding and Default cases represent values that should be predicated of persons themselves. I also think we aren't technically and conceptually equipped, at least for now, to empirically check whether an eluding or a default attitude occurs only in some modules or broadcast to a full suite of them.

CHAPTER 4:

GROUNDING MORAL RESPONSIBILITY ON ONE'S VALUINGS

Introduction

In chapter three, I suggested a valuing reading of the Expressivist Thesis and showed when we can ascribe someone a valuing. In this section, I argue that expanding the valuing reading of the Expressivist Thesis to the history of one's life describes what kind of morally responsible agents we are. I start by showing how expressivists' pictures of what it is to hold someone responsible also apply, implausibly, to non-blaming scenarios. To remedy this problem, I argue, neither the personal-level condition of evaluative judgments nor replacing evaluative judgment with valuing is sufficient. Rather, I argue, the blamed should be held responsible not only for her valuings but also for how she has come to them.

I. Grounding responsibility on valuing: negative cases

In chapter 3, I tried to show when subliminally instilled implicit attitudes can't show one's valuings. *A fortiori*, they can't change one's praise- or blameworthiness. But what Scanlon, Smith, and Arpaly suggest as the main thing involved in holding people responsible can be applied to these agents, especially when the effect of the manipulations lasts a long time, perhaps the whole life of the manipulated. Scanlon has to admit that we have reasons to avoid assigning these manipulated persons the task of judging about beverages or female

police officers. Similarly, we have reasons for adjusting our personal relationship when one's manipulated judgments are about us, like when subliminal manipulations have made our friend more sensitive to our faults and less appreciative of our positive points. For Arpaly, the faulty judgments of these people about female officers or our faults are like the poor aesthetic judgments of a lousy artist. Smith thinks that we are answerable for those mental "states 'in principle,' sensitive to our evaluative judgments [and not only for those] 'in fact' sensitive to these judgments" (2005: 253). A full evaluation of Smith's "in principle"-"in fact" distinction needs more space than what I have. I think it's at least unclear whether she can successfully exclude subliminally manipulated cases.

My valuing reading of expressivists excludes always-been sub-personal cases. Still, expressivists' history-insensitivity makes them treat other *not*-blameworthy cases as if they're blameworthy. Recall the Default cases. Some audience members of Janaan's talk may accept her claim very easily and commit themselves to it. This is a reason that their previous view has had no blameworthy root—like gaining the benefits of holding nonhuman animals inferior—and it may excuse them in Janaan's eyes. But according to Smith, Janaan has no other way than to hold them blameworthy because they have done many things and have had many attitudes *because of* their evaluative judgment about human superiority. Similarly, according to Arpaly, Jannan has to blame them because they have been, in Janaan's eyes, like a lousy artist. To exclude these not-blameworthy cases, some (moral-responsibility significant) role for one's coming to her valuings is needed.

II. Grounding responsibility on valuing: positive argument

I follow my valuing reading of expressivists to show what *kind* of role we can have for coming to our valuings. To understand how we really develop our valuings, two concerns are of special importance. First, as our valuings go beyond our consciously held propositional attitudes, our role in developing them can go beyond our conscious attempts too. Second, there can be no innocent self responsible for developing our valuings. Hence, I suggest that being a morally responsible agent is to be a valuing creature who develops her valuings through what I call the Responsibility Chain (RC) of her life:

> One's actions and attitudes have moral-responsibility significance only when they express one's current valuings—this is the valuing reading of the Expressivist Thesis. But all the elements of valuing can, conceptually, be implanted by a neurosurgeon. So to have moral-responsibility significance, all the dispositional and cognitive components of one's valuings should be, partly, up to oneself. This can be true only when one's current valuing elements are partly the result of some previous actions or choices expressive of one's valuings at the time they are made. The valuings at that time should, in turn, be partly the result of some previous actions or choices expressive one's valuings at the time they were made, and so on.[6]

The negative force of this simple argument arises from the fact that the chain goes *regressively* back to one's very young age when one isn't deserving of any blame or praise in the heaven-hell basic sense of desert that I discussed in chapter one. This negative conclusion isn't entirely novel for expressivists, but there is also a positive side of the RC Argument according to which our valuings are developed *progressively* throughout our lives and partly by *us,* where "us" means our previous valuings. This development makes room for the main intuition absent in expressivists' history-insensitivity, the intuition rooted in our formative experiences of deciding on some actions or choices that we guessed would influence the kind of persons we are going to become. This is how by our conscious choices and/or our unconscious habits we shape our future selves. For instance, we (should) care about habits that can make us lazy or diligent. This is a powerful intuition for the dependence of our blame- and praiseworthiness on ourselves. For example, we hold people like Nelson Mandela and Martin Luther King praiseworthy especially for their attempts to transform their hatred, cultivated from a long inequality, into a kind of love that includes people on all sides.

The RC Argument captures the history-sensitive intuition without denying the Beyond-Consciousness Thesis. Denying the Beyond-Consciousness Thesis necessitates a starting point for one's RC from which one decides what way of developing her valuings she likes

[6] The structure of this argument is like Galen Strawson's Basic Argument (Strawson, 1994). But, unlike him, the regress in my argument goes to our very young ages instead of showing and then rejecting the need for being *causa sui*. This difference gives my argument the positive side I discuss below.

to pursue. Despite the cartoonish nature of this picture, we can ask on what criteria is one supposed to base such a decision. For the same reason, assuming a beyond-RC self to whom credit or discredit can go to is incoherent or "unbearable" (Watson, 2016), independently and more importantly than being compatible or not with a general picture of the world as deterministic or indeterministic. So Mandela's and King's attempts to develop better things in themselves are the results of their already-cultivated valuings. Their already-cultivated valuings are, in turn, the result of their previous attempts/habits to develop better valuings in themselves. And so on.

III. How does my view explain the internal problems of responsibility?

In chapter two, I introduced the problem of blaming superbad people. The problem was rooted in the fact that evil people are less responsive to blame than most people. I also showed the possibility of the appropriateness of sadness for the case of superbad people as a possible threat for the deserved-anger view. In this section, I show how the RC view explains these apparent problems as some fundamental facts about our morally responsible agency.

Some philosophers, like Watson (2016) and Nelkin (2015) think that the fact that evil people are less responsive to moral reasons is a puzzle that needs a solution. The puzzle for these philosophers is a result of one of Watsonian Accountability's necessary conditions, according to which blame should be intelligible for the blamed, but "[i]n what sense can a deeply cruel man respond to reasons of kindness?" (Watson, 2016). The RC view, however, explains why this fact shows a fundamental feature of our moral-responsibility nature. We *shape* our valuings throughout our lives, and the more shaped our valuings, the less responsive they are to contrary reasons. Most people have complicated RCs and different, possibly conflicting, valuings that make them responsive to different signals about what is good or bad. They value furthering their self-interests and disvalue selfishness; they value the easiest possible paths to their goals and disvalue laziness; they value others' reactions and disvalue conventionality, and so on. However, because we are too focused on these normal people, narcissists, "bad apples", and Robert

Harris seem baffling to us. For the case of Harris, I think his difference is only that the RC of his life has a simpler narrative than many of us: Seeing his parents, who for everyone are supposed to be on their side, as evidently not so makes him see *no one* on his side and, as Watson diagnoses correctly, stand against the community as a whole. But Harris and "bad apples" share two fundamental, but mostly overlooked, features of our moral-responsibility nature, namely, first, decreased malleability as a consequence of more shaped valuings, and second, increased likelihood of adding more new bad (or good) valuings as a consequence of having already bad (or good) valuings. These features are not true only for evil people, as Watson and Nelkin worry. It's also true for good people. Mandela in his 50s is less concerned about others who might call him an unrealistic dreamer and more likely to extend his concerns to new issues like sexism, ageism, ableism, etc.

A metaphor for comparing the complexity of the RCs of normal lives and Harris' life may illuminate the RCs of human lives. Harris is like a simple mirror that reflects the influence of his parents. We can say that his awful present valuings *show* that what kind of person having terrible parents (when they see the child as unwanted and when there's no one to compensate for their effect) can lead to. More complex RCs, however, are like when small mirrors are put on a board at different angles that makes each mirror reflect a different thing: parents, teachers, books read, movies watched, and so on. To complete this metaphor for people with complex RCs, suppose that each new mirror is put on the board at any given time, partly by considering how it suits the composition of the other parts at that time. In this sense, the new pieces are partly *up to* the board itself. In this way, changes in our valuings are up to us, but there is no *us* who looks at the composition from an external point and decides for (non-)changes.

Knowing what kind of responsible agents we are, now we can see what kind of moral reactions can be, in principle, appropriate to us. In short, I think the practices of moral responsibility are the practices of co-valuing in which we value good and disvalue bad valuings *and* the RCs behind them. In blaming, we hold the blamed as a negative exemplar to tell her, ourselves, and/or others that this is how one should try not to be. More generally, the target of blame is one's valuings and the chain of previous valuings behind them, while the constitution of blame is the constitution of valuing: We see this target as disvaluable,

or at least as not valuable, and we are disposed both to respond to the reasons this disvaluing gives us for a variety of actions and attitudes and to feel some negative emotions against the target. Expressivists might be right that we don't need to look for details in one's background, but that's only because we assume enough complex RCs in the lives of most people unless we find contrary evidence, as in the case of Robert Harris.

IV. The "bearable" task of accounting for the responsibility for the self

At the end of section II, I mentioned how some philosophers like Watson think that the task of accounting for the responsibility for the self is unbearable. The view I advocate eludes many philosophers because of a presupposition about the options we have for understanding the idea of a person's having chosen to have certain characteristics. Scanlon, for instance, limits the options to two: 1) when the characteristic "comes about as a result of [one's] actual preferences and values" and 2) when the attitudes based on which someone chooses, in the first sense, to have a characteristic themselves chosen by the agent (Scanlon, 2008: 191). (1) is roughly what I described as the conditions of ascribing valuing to people. But (2) shouldn't be understood only as Scanlon suggests, namely being "independent of any factors that are not themselves ones the agent has chosen" (ibid.). Scanlon is right in seeing this understanding of (2) as "incoherent" (ibid.), and, following Nagel (1979: 35), as an idea that makes "the area of genuine agency, and therefore of legitimate moral judgment, [...] shrink [...] to an extensionless point" (Scanlon, 2008: 235n60). I've suggested a different interpretation of (2) as when one's current valuings are partly chosen by her, that is, to be chosen partly by her previous valuings. I hope I've explained why it's non-trivially more than (1) and non-trivially less than Scanlon's, among others, understanding of (2).

Another version of this presupposition thinks the Beyond-Consciousness Thesis is necessarily history-insensitive. Smith is explicit in saying that her problem with history-sensitive accounts (she calls them volitional accounts) is mostly because they deny, what I called the Beyond-Consciousness Thesis. She adds that she welcomes any future history-sensitive account that keeps the Beyond-Consciousness Thesis (2005: 265n38). A clear

instance of the presupposition in Arpaly is when she says that cases like Huck (and Oscar Schindler) “are obviously not praiseworthy for any kind of self-training or character-building on their parts. They are praiseworthy because, despite any character-building imposed on them by their misguided selves or others, some of their moral common sense, much of their moral goodness—that is, their responsiveness to moral reasons—remains intact” (80).

CONCLUSION

I defended a view of the morally responsible agent as a (co-)valuing creature. This view, I hope, describes actual human beings. Their (non-)actions and (non-)attitudes can show their valuings. Their valuings determine how they develop their later valuings. And they (dis)value good and bad valuings and the way they're developed both in themselves and in others.

Back to the sadness/resentment question from the introduction, this view provides two reasons for moral sadness to the extent one is blameworthy and for counting oneself morally lucky to the extent one is praiseworthy: 1) We are the results of nothing beyond constitutive luck, represented as the RCs of our lives and we don't want an RC that leads to bad valuings. 2) Being (more) blameworthy means being less (valuingly) disposed to the reasons for (moral) improvement either by being blamed or by other things which might make the badness of one's valuings evident to her. Similarly, being (more) praiseworthy means having a lesser tendency to get worse by either the criticisms of others or other things that might make good valuings hard. At the same time, being more blameworthy means being more strongly in need of seeing the signs of bad valuings. For instance, the more a person values friendship only for short-term joys or benefits, the more she loses the goodness of deep friendship. But it means also, sadly, the less she is responsive to the signs that invite her to value deeper friendships. As I argued in chapter 2, some personality disorders, most saliently narcissism, are nothing beyond being too bad. If my point here is right, this being too bad in itself is a reason for holding a person an apt target of (moral) sadness.

I'm not sure what I can say about resentment, because knowing what kind of responsible agents we are is only one side of its story. The other side is its nature. However, I can say that *if*, and only if, resentment can target only one's valuings and the RC behind it, and not a sense of self beyond it, it can be appropriate. The litmus test is that resentment should be compatible, and even correlate, with moral sadness.

I finish with what this view implies for punishment. Even people with the worst valuings and complex enough RCs behind them aren't deserving of any extra pain that is not either the necessary consequence of their valuings, like being not a party of deep friendships for one who values friendship only for short-term joys or benefits, or done for either consequentialist reasons, like moral improvement, or prudential reasons, like protecting others from their dangers. As a specific instance, we shouldn't think that even war criminals don't deserve to have good views in their cells independent of such consequentialist reasons.

According to the quality-of-valuing view I defended in this thesis, the deserved-angry blame seems functional in the cases of morally mediocre people because: 1) they are disposed to many bad reasons and their contrary good reasons in many cases. 2) they are disposed to the reasons for avoiding (self-)blame. Unlike morally mediocre people, a narcissist is someone who needs professional help not because he is exempted, but because of the very fact that he has become too bad to be able to be improved by being blamed. As I argued, this claim should not be put in the exempted/responsible dichotomy. Rather, my claim is that our natural framework of moral responsibility practices is not strong enough for understanding superbad people. I suggested the basic lines of how to understand them. Based on my view, I think we have a long way ahead for finding the appropriate moral reactions to superbad people.

REFERENCES

American Psychiatric Association, (1996). *DSM-IV-TR*.

——. (2013). *DSM-5*.

Arpaly, N. (2002). *Unprincipled virtue: An inquiry into moral agency*. OUP.

——. (2005). How it is not" just like diabetes": Mental disorders and the moral psychologist. *Philosophical Issues*, 15, 282-298.

Callard, A. (2017). The reason to be angry forever. *The moral psychology of anger*, 123-137.

Elton, M. (2000) The personal/sub-personal distinction: An introduction. *Philosophical Explorations*, 3(1), 2-5.

Fischer, J. M. (2006). The cards that are dealt you. *The Journal of Ethics*, 10(1-2), 107-129.

Kozuch, B., & McKenna, M. (2015). Free Will, Moral Responsibility, and Mental Illness. In *Philosophy and Psychiatry: Problems, Intersections and New Perspectives*. Taylor and Francis Inc.

Jaworska, A. (2016). Holding psychopaths responsible and the guise of the good. *Current controversies in bioethics*, 66-77

Levy, N. (2007). The responsibility of the psychopath revisited. *Philosophy, Psychiatry, & Psychology*, 14(2), 129-138

——. (2011). *Hard luck: How luck undermines free will and moral responsibility*. OUP.

——. (2014). *Consciousness and moral responsibility*. OUP.

Mele, A. (1995). *Autonomous agents*. OUP.

Nagel, T. (1979). Mortal questions. CUP.

Nelkin, D. K. (2015). Psychopaths, incorrigible racists, and the faces of responsibility. *Ethics*, 125(2), 357-390

Pickard, H. (2011). Responsibility without blame: Empathy and the effective treatment of personality disorder. *Philosophy, psychiatry, & psychology*, 18(3), 209.

——. (2013). Responsibility without blame: philosophical reflections on clinical practice. *Oxford handbook of philosophy of psychiatry*, 1134-1154.

Pereboom, D. (2013). Free will skepticism, blame, and obligation. In *Blame: Its nature and norms*, 189-206.

Pereboom, D. (2014). *Free will, agency, and meaning in life*. OUP.

Scanlon, T. M. (2008). Moral dimensions: Permissibility, meaning, and blame. HUP.

——. (2015). Forms and conditions of responsibility. In *The Nature of moral responsibility*. Eds. Randolph Clarke, Michael McKenna, and Angela Smith. OUP, 89-111.

Scheffler, S. (2011). Valuing. In *Reasons and recognition: Essays on the philosophy of T. M. Scanlon*. OUP, 23-42.

Smith, A. M. (2005). Responsibility for attitudes: Activity and passivity in mental life. *Ethics*, *115*(2), 236-271.

——. (2015). Responsibility as answerability. *Inquiry*, 58(2), 99-126.

Sripada, C. (2015). Commentary on Kozuch and McKenna: Mental Illness, Moral Responsibility, and Expression of the Self. In *Philosophy and Psychiatry: Problems, Intersections and New Perspectives*.

Strawson, G. (1994). The Impossibility of moral responsibility. *Philosophical Studies*, 75(1–2), 5–24.

Strawson, P. F. (1982 [1962]). Freedom and resentment, reprinted in G. Watson (Ed.), *Free will*.

Strohminger, N. (2014). Disgust talked about. *Philosophy Compass*, 9(7), 478-493.

Watson, G. (2011). The trouble with psychopaths. In *Reasons and recognition: Essays on the philosophy of T. M. Scanlon*. OUP,

——. (2016). Responsibility and the limits of evil: Variations on a Strawsonian theme. In *Free will and reactive attitudes*. Routledge, 127-154. Reprinted with an added foreword from F. Schoeman (ed.), *Responsibility, Character, and the Emotions* (1987), CUP, pp. 256–86.

Wolf, S. (2011). Blame, Italian Style. In *Reasons and recognition: Essays on the philosophy of T. M. Scanlon*. OUP.