

**From the holdings of the Department of Special Collections , Princeton University Library,
Princeton, NJ**

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyright law. We do not charge any permission or usage fees for the publication of images of material in our collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff. For more information regarding copyright, credit, and citation please review the information on the Copyright, Credit and Citation Guidelines page of our website.

<https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines>. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. <https://library.princeton.edu/special-collections/ask-us>

Doctoral Dissertations, 2011-2026

Philosophy

2016

A Theory of Blame and Blameworthiness

Gease, Arlyss

Princeton, NJ : Princeton University

<http://arks.princeton.edu/ark:/88435/dsp01pg15bh37q>

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. <https://www.copyright.gov/title17/title17.pdf>

A THEORY OF BLAME AND BLAMEWORTHINESS

Arlyss Logan Gease

A DISSERTATION

PRESENTED TO THE FACULTY

OF PRINCETON UNIVERSITY

IN CANDIDACY FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE

BY THE DEPARTMENT OF

PHILOSOPHY

Advisers: Gideon Rosen, Elizabeth Harman

November 2016

© Copyright by Arlyss Logan Gease, 2016. All rights reserved.

Abstract

This dissertation develops a theory of moral blame and of blameworthiness, the conditions under which an agent is an appropriate target for moral blame.

Chapter One argues against the thesis that the facts that determine whether an agent is blameworthy for some behavior are facts to which she stands in a certain epistemic relationship. It is true that an agent is blameworthy for the consequences of some act or omission only if she knows of the risk that her behavior will have such consequences. However, an agent can be blameworthy because she acts on certain morally objectionable motivations even if she does not stand in any particular epistemic relationship to the fact that she is so motivated. What ultimately matters to blameworthiness is not an agent's epistemic state, but the quality of will expressed by her behavior.

Chapter Two gives a precise account of when an agent expresses the kind of morally objectionable quality of will that is necessary for blameworthiness. I argue that such a quality of will is expressed when (i) an agent's beliefs and desires rationalize behavior that (ii) the agent possesses moral reasons not to engage in, reasons that (iii) the agent's motivations fail to counterbalance. The reason that blame is governed by these conditions, I argue, is that blaming someone involves having a thought about the offending agent's motivations that is true only if these conditions are met.

Chapter Three argues that any moral norm that bears on whether an agent merits sanctions for her behavior also bears on whether she is blameworthy for that behavior. The chapter answers recent criticism of this position by identifying a feature of moral blame that explains why blame should be governed by the same moral norms that determine whether an agent merits

sanctions, namely, that in blaming someone for engaging in some behavior, one takes there to be no backward-looking moral principles that prohibit sanctioning her for that behavior. This discussion motivates a novel account of blame that contrasts with pure cognitivist theories of blame and theories on which blame is essentially retributive.

Acknowledgments

I have a great many people to thank for contributing to this dissertation. First amongst these is Yusha Hu, who has been my greatest source of support throughout.

Also remarkable is the introduction to philosophy I received at Swarthmore College. As an undergraduate major in both philosophy and mathematics, my interests were far removed from the topic of this dissertation, but what abilities I have to engage with challenging problems in philosophy owe a great deal to my instructors at Swarthmore, Peter Baumann and Richard Eldridge in particular. I am especially grateful to Alan Baker, who advised my undergraduate thesis and truly prepared me to pursue a PhD.

After joining the Princeton University Department of Philosophy and beginning to work in earnest on topics related to blame and responsibility, I benefitted a great deal from the Moral Responsibility Reading Group that met in the spring of 2013. Special thanks to Daniel Wodak, Benjamin Ewing, Angela Smith, and Peter Graham, who contributed greatly to my thinking about this topic. Thanks also to my colleagues Hrishikesh Joshi, Adam Lerner, and Sam Preston, who made helpful contributions to earlier drafts of this material, and to Michael Smith, who has been an important influence on this work.

No one has contributed more to this project than my dissertation advisers, whose incredible generosity I can only begin to explain. The quality of my philosophical work has never improved more rapidly than in the months after Elizabeth Harman became one of my advisers on this project. This dissertation is so much better for your feedback and encouragement. Finally, I am so grateful to Gideon Rosen. For inspiring the confidence to begin this project and the commitment to see it through, thank you.

Contents

Introduction	1
1. Epistemic Conditions on Blameworthiness	12
1.1 Blameworthiness for Consequences	14
1.2 General Epistemic Conditions	24
1.3 Objections	31
2. Quality of Will	47
2.1 Moral Reasons and Rational Control	49
2.2 Quality of Will and the Nature of Blame	65
2.3 Objections	71
3. Blameworthiness, Sanctions, and Desert	81
3.1 Sanctions and the Nature of Blame	84
3.2 Blameworthiness and Desert	94
3.3 Objections	102
Conclusion	110
References	118

Introduction

This is an essay about moral blame. One blames an agent for some behavior, in virtue of the morally objectionable attitude one takes that behavior to express. Following Strawson [1962], I take the paradigm forms of blame to be moral resentment, indignation, and guilt.¹

Attitudes of moral blame are important both to interpersonal relationships and to the law. Blame affects our behavior, most obviously by motivating us to express moral criticism — whether in words or unwelcome social sanctions — and in subtler ways as well. Blaming practices, along with related practices of apology and forgiveness, are an essential part of ordinary moral life. As to the law, it is part of the function of criminal punishments to give voice to an attitude of moral blame or condemnation.² All of this makes blame an important subject for moral philosophy.

Like criminal punishments, which are unjust when imposed on a person who has not committed a crime — or, just as seriously, when imposed on a person for behavior that should not be a crime — moral blame is inappropriate when directed at an agent who has not engaged in the relevant kind of morally objectionable behavior. Such an agent is not *blameworthy*, not an appropriate target of moral blame. Consider a case.

Shiitake: Suzy is making a shiitake mushroom broth to have for dinner with Tiffany.

Unbeknownst to Suzy, Tiffany is dangerously allergic to shiitake. Suzy has never heard

¹ Peter F. Strawson, “Freedom and Resentment”, *Proceedings of the British Academy*, 1962.

² Feinberg [1965] argues persuasively for this point. (Joel Feinberg, “The Expressive Function of Punishment”, *The Monist*, 1965)

of a shiitake allergy, nor does Tiffany inquire into the provenance of the broth before beginning to eat. As a result, Tiffany suffers a severe allergic reaction.

Suzy is not blameworthy for causing Tiffany to have a severe allergic reaction, even though causing someone to have such an allergic reaction is the kind of thing for which one might be seriously blameworthy. Suzy is blameless because there is some condition of blameworthiness, some condition that must be satisfied in order for an agent to be blameworthy for her behavior, that Suzy does not satisfy. What are the conditions of blameworthiness? And why is the appropriateness of blame governed by these conditions? These are the questions motivating the three chapters of this essay.

Epistemic Conditions

In Chapter One, I discuss the *epistemic conditions* on blameworthiness, conditions that have to do with an agent's knowledge, beliefs, and evidence. An epistemic condition on blameworthiness promises to explain cases like Shiitake, since Suzy's ignorance of Tiffany's allergy is clearly relevant to why she is blameless for the bad outcome of her behavior.

I begin my discussion of epistemic conditions with a puzzle. Even supposing that an agent is blameworthy for some act or omission, it can happen that the agent is to blame for some of the consequences of her behavior but not others. Consider a case.

Meeting: Mary receives a call from Nina, who asks Mary to meet her socially at a certain cafe at a certain time. Mary promises Nina that she will be there. Unbeknownst to Mary, there is a spy, Olivia, who plans to assassinate Nina. Nina has asked Mary to be present at the cafe to discourage Olivia from assassinating her.

When the time comes for Mary to go to the cafe, she is watching an entertaining television show. Although Mary is usually very diligent about keeping her appointments, she decides to finish the show before leaving to meet Nina. As a result, Mary arrives at the cafe twenty minutes late. In that time, Olivia kills Nina. Had Mary been on time, Nina would not have been killed.

One can blame Mary for the decision to finish watching the television show. One can blame Mary for arriving at the cafe twenty minutes late, a consequence of that earlier decision. One should not, however, blame Mary for Nina's death. The explanation concerns Mary's epistemic state, namely, the fact that Mary does not know that arriving late creates a risk that Nina will be killed. I argue that Meeting and other cases support the following principle.

Knowledge of Risk: An agent A is blameworthy for some consequence C of behavior X only if

- (i) A believes on the basis of evidence E that X might result in a consequence of kind K, and
- (ii) C realizes K in a way such that E is not completely defeated.

That blameworthiness for consequences is governed by the conditions given in Knowledge of Risk is an important result. However, no principle about blameworthiness for consequences can fully explain why an agent like Suzy is blameless in a case like Shiitake. Suzy's ignorance of Tiffany's allergy is relevant not only to whether she is blameworthy for any symptoms that Tiffany suffers, but is also relevant to whether she is blameworthy for the decision to serve shiitake in the first place. To fully explain cases like Suzy's, many philosophers accept a more general principle like the following.

General Explanatory Condition: The fact that an agent A is blameworthy for behavior X is sufficiently explained by facts of the form E(A, P), where E is an epistemic attitude that A bears towards some morally significant proposition P.

For example, some philosophers believe that agents are blameworthy only for behavior that they know to be wrong, because they know it to be wrong. Although it is wrong to risk causing someone to have a severe allergic reaction, because Suzy doesn't know that she is doing anything wrong, she is blameless. Or so the story goes.

On my view, however, General Explanatory Condition is false. A sufficient explanation of why an agent is blameworthy sometimes depends on facts to which the agent has no epistemic access. Consider the following case.

Interview: Irene is interviewing Janet for a position at a company. At one point, Janet demonstrates her qualification for the position by telling a story that, unbeknownst to Janet, vividly recalls for Irene some intensely embarrassing memories. As a result, although Irene is sensitive to Janet's strengths as a candidate, she leaves the interview with an unfavorable impression of Janet. Irene does not have the self-awareness to realize that her personal feelings might have an influence on how she evaluates Janet's candidacy.

When Irene is later asked to evaluate Janet's suitability for the position as part of the hiring committee's group deliberation, Irene testifies that Janet may not perform well on the job and gives some inconclusive evidence to this effect. While Irene would not draw attention to this evidence unless it were relevant to the hiring group's deliberation, Irene

is not motivated by a desire to hire the best candidate; her true motivation for calling attention to this evidence is simple dislike of Janet. However, Irene has no idea that this is what is motivating her to testify against Janet, nor could it occur to her.

I claim that Irene's case is a counterexample to General Explanatory Condition because (i) Irene is blameworthy for the decision to testify against Janet in virtue of the fact that she is motivated by dislike, but (ii) Irene does not stand in any positive epistemic relationship to her true motives for testifying against Janet.

Why doesn't Irene's ignorance about her subconscious motivations for testifying against Janet render her blameless for her biased testimony the same way that Suzy's ignorance of Tiffany's shiitake allergy renders Suzy blameless for serving Tiffany shiitake broth? The answer, I argue, is that an agent's epistemic state is relevant to blameworthiness in virtue of a more fundamental condition of blameworthiness, the condition that an agent's behavior express a *morally objectionable quality of will*, a lack of concern for the rights and interests of others.³ An agent's epistemic state contributes to the quality of will expressed by her behavior, and it determines which of the consequences of an agent's behavior express that will; however, an agent's epistemic state is not the only factor that bears on the quality of will expressed by her behavior. On my view, the fact that Irene is motivated to testify against Janet out of dislike makes her testimony express a morally objectionable quality of will even though Irene is not aware of her motivations. Suzy is blameless in Shiitake, on the other hand, because her motivations for serving Tiffany the broth involve no morally objectionable quality of will.

³ The language of "quality of will" comes from Strawson [1962].

Quality of Will

Chapter Two focuses on the quality of will condition on blameworthiness. The ambition of this chapter is to both identify what precisely is required for an agent to express a morally objectionable quality of will and, further, to explain why blameworthiness requires that this condition be satisfied.

An acceptable account of when an agent's behavior expresses a morally objectionable quality of will must distinguish behavior that expresses an agent's will from behavior that does not.

Consider a case.

Kitten: Kate is across the street from a burning building where firefighters are trying to rescue a trapped kitten. They do not know what room the kitten is in and the fire is spreading fast. Soon the firefighters will be forced to exit the building. Kate is extremely concerned for the kitten and wishes she could help.

Suddenly, Kate sees movement in a window and realizes that it's the kitten. Kate could call out to the firefighters and the kitten would be saved. All things considered, this is what Kate desires to do. However, Kate finds herself frozen with panic. A moment later, Kate loses her opportunity to act when the fire chief calls an exit from the building.

Normally, an agent who fails to take action to rescue a kitten from a fire would be blameworthy for her inaction. However, although Kate could and should yell out, her failure to do so does not express a morally objectionable quality of will. It does not express her will at all.

Intuitively, what is important about Kate's case is that she does not fail to rescue the kitten because her desire to rescue the kitten is too weak or because she has stronger competing desires; rather, Kate fails to act because she is paralyzed by irrational fear. In the chapter, I argue that only when an agent's behavior is *rationalized* by her preferences — that is, only when it is licensed by the agent's capacity for means-end reasoning, given what she believes and desires — does that behavior express her will. Kate's inaction in Kitten does not express her will because such inaction is not rationalized by her preferences, since her preferences would be best served by rescuing the kitten.

Assuming that an agent's behavior is rationalized by her preferences, when does her behavior express a will that is morally objectionable in the sense required for moral blame to be appropriate? The quality of will condition that I endorse is in the vein of proposals made by Arpaly [2002] and Markovits [2010] according to which the quality of will expressed by an agent's behavior depends on how the agent's motivations compare to the moral considerations that ought to be motivating her.⁴ I argue that an agent expresses a morally objectionable quality of will when her preferences rationalize behaving as she does, and when her motivations for so acting fail to counterbalance the moral reasons she has to act otherwise. For instance, when Irene testifies against Janet in Interview, Irene expresses a morally objectionable quality of will because her behavior is rationalized by her dislike of Janet, and because she is not motivated by the kind of considerations that normally justify such negative testimony (concern over whether Janet will perform well on the job, contribute to the company, etc.). In the chapter, I show how this explanation applies in a wide range of cases.

⁴ Nomy Arpaly, *Unprincipled Virtue*, Oxford University Press, 2002.
Julia Markovits, "Acting for the Right Reasons", *The Philosophical Review*, 2010.

Why is it that the appropriateness of moral blame depends on an agent's expressing, in this sense, a morally objectionable quality of will? The answer, I believe, concerns the kind of moral criticism that is involved in moral blame. Part of what distinguishes resentment and indignation from other forms of anger, and distinguishes guilt from other forms of regret, I argue, is that these attitudes all involve a thought about the moral quality of the offending agent's motivations. This thought is a *cognitive state*, a mental state that presents things as being a certain way and which could be true or false. It is relatively uncontroversial that an attitude is appropriate only if its cognitive elements are true.

Cognitive Conditions: If an attitude A involves a cognitive element with content P, then A is appropriate only if P is true.

The thought about the offending agent's motivations that is involved in a particular instance of resentment, indignation, or guilt varies from case to case — one might think that the agent's behavior was callous, for instance, or malicious, or that the agent should have taken some earlier precaution against some known risk — but in every case, the attitude is appropriate only if the thoughts it involves are true. The reason that there is a quality of will condition on moral blameworthiness, on my view, is that part of what makes an attitude a form of blame is that it involves thoughts that entail that the offending agent has expressed a morally objectionable quality of will. It follows from Cognitive Conditions that blame is only ever appropriate when directed at an agent who has in fact expressed a morally objectionable quality of will.

Sanctions

In the first two chapters of this essay, I trace the significance of one fact about an agent that bears on blameworthiness — her epistemic state — to a more fundamental condition on

blameworthiness, the quality of will condition, which holds that an agent is blameworthy only for behavior that expresses a morally objectionable quality of will. This condition is fundamental, on my view, because it is directly explained by the thoughts involved in blame.

Are we finished? Is this a complete theory of moral blame and blameworthiness? Not according to the view I develop in Chapter Three. The arguments in the first two chapters so far do not explain why moral blame is naturally expressed in the form of moral sanctions, including harsh moral criticism and other forms of treatment that symbolize such criticism. Chapter Three gives an account of this relationship and how it contributes to the conditions of blameworthiness.

Sanctions are an important subject for moral philosophy. Morality generally prohibits treating others in ways that are unwelcome or unpleasant, but there are exceptions. For instance, in cases where an agent knowingly and without defense engages in behavior that violates a just criminal statute, moral sanctions are generally permissible. On my view, the reason that moral blame is naturally expressed in the form of sanctions is that blaming someone involves taking there to be no moral principle of a certain kind that prohibits sanctioning her for the behavior in question.

Taking a fact to bear on what one ought to do — for instance, to bear on whether one should sanction someone — is a *conative state*, a practical state that affects an agent's behavior. And much like cognitive states, I argue, which make the appropriateness of an attitude depend on the truth of some proposition, conative states make the appropriateness of an attitude depend on certain norms.

Conative Conditions: If an attitude A involves taking a fact F to bear on whether one should Φ in virtue of norm N, then A is appropriate only if F does bear on whether one should Φ in virtue of norm N.

In the case of blame, the relevant norms are *backward-looking moral prohibitions on the imposition of sanctions*. These are norms that prohibit imposing sanctions on agents for certain kinds of behavior or for behavior that was performed in a certain way. On my view, blaming someone involves taking her behavior (F) to bear on whether one should sanction her (Φ) in virtue of being the kind of behavior for which backward-looking moral principles do not prohibit sanctions (N). According to Conative Conditions, then, blame is appropriate only if there truly are no backward-looking moral norms that prohibit sanctioning the agent for her behavior.

On this view, the way that blame tends to dispose one to sanction an offending agent is by removing a barrier to imposing sanctions so that, if the evidence suggests that imposing sanctions would be good in some way, one has license to impose them. It is noteworthy that this view does not imply that blame always involves taking there to be reasons that count in favor of going ahead with sanctions. This is as it should be: One can blame someone without having a desire that she suffer sanctions. In the case of guilt in particular, an agent might feel that it would not be unfair for others to sanction her, but have every hope that no sanctions will be imposed.

In Chapter Three, I suggest that the two features of blame identified so far in this essay — first, a thought about quality of will, second, taking an agent's behavior to license sanctions — allow us to give a precise definition of the attitudes that constitute moral blame.

Blame Attitudes: For an agent A to blame an agent B for behavior X is for A to have an attitude E that involves:

- (i) thinking that B's performance of X has features that, as a matter of fact, are sufficient for X to express a morally objectionable quality of will, and
- (ii) taking there to be no backward-looking moral principle that prohibits sanctioning A for X.

This account implies that there are just two fundamental conditions of blameworthiness. The thought in (i) makes blameworthiness depend on an agent's expressing a morally objectionable quality of will. The conative attitude in (ii) makes blameworthiness depend on an agent's behavior lifting substantive constraints on the imposition of sanctions. The view thus captures both the sense in which moral blame involves a form of deep moral criticism and also explains the relationship between blame and moral sanctions.

Both conditions imply that whether a particular agent is blameworthy for a particular act depends not only on the nature of moral blame, but on substantive moral norms, namely, on backward-looking moral prohibitions on the imposition of sanctions and on whether some motive morally justifies an agent's behavior, given her circumstances. This essay does not aim to resolve such questions, but rather to show how quite ordinary moral attitudes relate to these substantive issues in moral philosophy.

1. Epistemic Conditions on Blameworthiness

This chapter argues against the thesis that the facts that determine whether an agent is blameworthy for some behavior are facts to which she stands in a certain epistemic relationship. It is true that an agent is blameworthy for the consequences of some act or omission only if she knows of the risk that her behavior will have such consequences. However, an agent can be blameworthy because she acts on certain morally objectionable motivations even if she does not stand in any particular epistemic relationship to the fact that she is so motivated. What ultimately matters to blameworthiness is not an agent's epistemic state, but the quality of will expressed by her behavior.

It is widely held that there is an epistemic condition on blameworthiness, that is, that to be blameworthy for engaging in some morally objectionable behavior, an agent must have some kind of relevant evidence, knowledge, or belief about that behavior's morally relevant features.⁵ Epistemic conditions on blameworthiness account for cases like this one.

Shiitake: Suzy is making a shiitake mushroom broth to have for dinner with Tiffany.

Unbeknownst to Suzy, Tiffany is dangerously allergic to shiitake. Suzy has never heard of a shiitake allergy, nor does Tiffany inquire into the provenance of the broth before beginning to eat. As a result, Tiffany suffers a severe allergic reaction.

⁵ By "behavior" I mean any of an agent's mental states or bodily movements, or causal chains beginning with an agent's mental states or bodily movements. If an agent cannot be blameworthy for some subset of her behavior — for instance, for involuntary tics or unforeseeable consequences — this is a substantive thesis that requires explanation.

Suzy causes Tiffany to suffer an allergic reaction. Objectively, this is a terrible thing to do, but Suzy is not blameworthy. The standard explanation for why an agent like Suzy is not blameworthy is that Suzy has no reason to suspect that Tiffany is allergic to the broth, and that this fact about Suzy's epistemic state — her ignorance of the risk she is creating — constitutes an *excuse*, that is, that Suzy's epistemic state shows that she does not satisfy some condition that an agent must satisfy in order to be blameworthy for her behavior.⁶

In the first section of this chapter, I focus on the epistemic condition on blameworthiness for the consequences of blameworthy behavior.⁷ The question of how blameworthiness for some behavior extends to the downstream consequences of that behavior is an issue that has received relatively little attention in the philosophical literature.⁸ When philosophers do discuss the issue, they are generally content to say that an agent is blameworthy for some bad outcome because it was “foreseen” or “foreseeable” by the agent. I offer and defend a more precise account on which an agent is blameworthy for some bad outcome of her behavior only if the outcome realizes a risk that was known to the agent when she acted.

In the second section of the chapter, I turn to consider epistemic conditions on blameworthy behavior more generally. I argue against the thesis that agents always bear some epistemic attitude to the facts that make them blameworthy. It is possible, I claim, for an agent to be blameworthy in virtue of facts of which she is totally unaware. In particular, an agent's motivations can make her blameworthy even if she is completely unaware that she has those

⁶ Austin [1956] popularized the notion of excusing conditions. (J.L. Austin, “A Plea for Excuses”, *Proceeding of the Aristotelian Society*, 1956)

⁷ We can assume that blameworthy consequences are always the consequence of a blameworthy act or omission since causing or failing to prevent some bad outcome is itself an act or omission, and any agent who is blameworthy for an outcome is necessarily blameworthy for causing (or failing to prevent) that outcome.

⁸ It should be noted that Feinberg [1970] and Thomson [1986] provide excellent discussions on related topics. (Joel Feinberg, *Doing and Deserving*, Princeton University Press, 1970; Judith Jarvis Thomson, *Rights, Restitution, and Risk*, Harvard University Press, 1986)

motives. This fact is explained, I argue, by the nature of moral blame, the fact that moral blame is a reaction to the moral quality of will expressed in an agent's behavior. An agent needs to stand in some epistemic relationship to a possible outcome of her behavior in order for that outcome to express her will. An agent's motivations, on the other hand, necessarily express her will, even if she is not aware of them, because an agent's motivations constitute her will.

In the third section, I consider some objections to the views defended in the first two sections of the chapter. I consider an objection that holds that we have not accounted for all of the ways in which an agent's epistemic state affects blameworthiness, objections that hold that the view is either too restrictive or too liberal in assigning blame for consequences, and an objection that agents can be blameworthy for negligent actions that involve ignorance of the relevant risks. These are by no means the only possible objections to the view, but each identifies important consequences of the view or issues worthy of further consideration.

1.1 Blameworthiness for Consequences

It is possible for an agent to be blameworthy for engaging in some behavior and to yet be blameless for some of that behavior's bad consequences. Consider a case.

Meeting: Mary receives a call from Nina, who asks Mary to meet her socially at a certain cafe at a certain time. Mary promises Nina that she will be there. Unbeknownst to Mary, there is a spy, Olivia, who plans to assassinate Nina. Nina has asked Mary to be present at the cafe to discourage Olivia from assassinating her.

When the time comes for Mary to go to the cafe, she is watching an entertaining television show. Although Mary is usually very diligent about keeping her appointments,

she decides to finish the show before leaving to meet Nina. As a result, Mary arrives at the cafe twenty minutes late. In that time, Olivia kills Nina. Had Mary been on time, Nina would not have been killed.

Mary is blameworthy for the decision to finish watching the television show. She is also blameworthy for arriving late to her meeting with Nina, a consequence of her prior blameworthy decision. But although Nina dies as the result of Mary's actions, Mary is not blameworthy for Nina's death.

Intuitively, the explanation for why Mary is blameworthy for some of the consequences of her actions but not others concerns the risks that are known to Mary when she is making the decision to finish watching the show. She knows of the risk that she will arrive late to her meeting with Nina, breaking her promise. She does not know of the risk that Nina will be killed.

On reflection, however, an explanation in terms of "known risks" is itself somewhat puzzling, invoking as it does both the concept of knowledge, a privileged epistemic state that is necessarily true, and the concept of risk, which goes hand in hand with uncertainty. A simpler view holds that the required state is not knowledge, but just belief. Mary is blameless for Nina's death, on this view, because she does not believe that finishing the show risks causing Nina's death. However, blameworthiness for consequences requires more than just belief. Consider another case.

Concrete: Carla is laying a foundation for a house and in the course of some calculations comes to the incredible conclusion that, given the slope of the land, the soil conditions, and various other features of the project, without reinforcement, an ordinary concrete foundation will become unstable and the house will collapse. Carla is initially

shocked by her findings, then realizes that this is the perfect opportunity to kill her despised sister, Dolores, who will occupy the house. Intent on killing Dolores, Carla does not reinforce the foundation.

In fact, Carla is right to find it incredible that the house will collapse unless the foundation is given additional reinforcements: Carla's conclusion was the result of a gross error in her calculations, and no evidence she has about the project conditions makes the conclusion even remotely plausible.

However, as the result of a dramatic shift in weather patterns, a gradual rise in the water table, changes in local soil composition, and erosion, the foundation of Dolores's house becomes unstable and the house collapses, killing Dolores. Had Carla reinforced the foundation, the house would not have collapsed.

Carla is plausibly blameworthy for attempting to kill Dolores, but she is not blameworthy for Dolores's death. This is so even though Carla's belief that Dolores will be crushed in the collapsing house is true. Why then isn't she blameworthy for Dolores's death? Because despite Carla's belief that Dolores will be crushed, the event itself is a kind of freak accident, totally unsupported by the available evidence. Concrete shows that an agent is only blameworthy for some bad outcome of her behavior if she is appropriately sensitive to a real risk of that outcome.

Concrete shows that blameworthiness for the consequence of an action depends not only on an agent's beliefs, but also on the evidence that supports those beliefs. However, it is important to

allow that even relatively weak evidence of a risk can make an agent blameworthy. Consider a case from Rosen [2008].⁹

Dorfman poisons Mrs. Dorfman by putting what he takes to be arsenic in her tea. The stuff is indeed arsenic and Mrs. Dorfman dies as planned. But Dorfman does not know that the stuff is arsenic (or that his act subjects his victim to an unjustifiable risk of death) because:

The chemist who sold Dorfman the arsenic is a famous liar. As he handed over the glassine envelope, he assured Dorfman that it contained a lethal dose of arsenic, and Dorfman believed him. But given the chemist's well-known track record of selling sugar as arsenic to would-be poisoners, Dorfman had no business believing him. Dorfman's pertinent beliefs are true, but they do not amount to knowledge because they are based on insufficient evidence. (Rosen [2008], 596)

Rosen claims that Dorfman's belief that the stuff in the envelope is arsenic makes him blameworthy for Mrs. Dorfman's death. Cases like Concrete, however, show that this explanation is incomplete. On my view, Dorfman is blameworthy because he *possesses evidence* that his actions will cause Mrs. Dorfman's death; even though the chemist is a famous liar, the facts of which Dorfman is aware still support raising his credence that he is poisoning his wife. This is why, if Dorfman were to come to his senses and realize that the stuff the chemist sold him is probably sugar, he would still have no business putting it in Mrs. Dorfman's tea.

⁹ Gideon Rosen, "Kleinbart the Oblivious and Other Tales of Ignorance and Responsibility", *The Philosophical Review*, 2008.

One might at this point wonder whether we have identified a meaningful difference between Dorfman and Carla from Concrete. Surely, one might think, the fact that Carla's calculations indicate that the house will collapse gives her at least some evidence that the house will indeed collapse; so long as Carla does not know that her calculations are mistaken, she should raise her credence that the house will collapse at least a little. Given that the unreliability of Dorfman's chemist does not prevent him from being blameworthy for killing Mrs. Dorfman, one might be tempted to reconsider whether Carla is blameworthy for killing Delores.

The crucial difference between Carla and Dorfman, however, is not that Dorfman possesses evidence that supports increasing his credence that he is causing Mrs. Dorfman's death whereas Carla does not possess evidence that supports increasing her credence that she is causing Delores's death. Rather, the difference concerns the relationship between these agents' respective evidence and the further facts of their circumstances. Carla's evidence that Delores will be crushed by the collapsing house — the apparent result of her calculations — is completely defeated by the fact that her calculation involves a gross error and that nothing she knows about the project conditions supports this conclusion.¹⁰ On the other hand, Dorfman's evidence that Mrs. Dorfman will die — the fact that the chemist says the stuff in the envelope is arsenic — is only weakened, but not fully defeated by the fact that the chemist is a famous liar. On my view, this is how to understand what it means to say that an agent "knows of a risk" in

¹⁰ I here assume the view that, like all epistemic reasons, *epistemic defeaters* — considerations that undermine or rebut other epistemic reasons to hold some belief — are facts. Sometimes these are facts about an agent's epistemic state: The fact that all of the evidence of which Carla is aware supports the conclusion that Delores's house does not need a reinforced foundation is a fact about Carla's epistemic state that defeats her reasons for believing that if she does not reinforce the foundation, the house will collapse. Other times, however, defeaters are facts of which an agent is unaware. If Brittany fills a bottle with water, which Clarissa then secretly replaces with gin, which Dianne secretly replaces with water, Brittany's reason for believing that the stuff in her bottle is water — namely, that she filled it with water to begin with — has been undermined by Clarissa and Dianne's subsequent behavior; this is why her true belief that the stuff in her bottle is water does not amount to knowledge, despite being justified relative to the evidence Brittany possesses. I consider cases similar to this one in section 1.3, which provide further support for thinking that facts outside of an agent's epistemic possession can defeat her evidence.

the sense relevant to blame: Dorfman knows his behavior risks harming Mrs. Dorfman, whereas Carla does not know her behavior risks harming Delores, because only Dorfman appreciates evidence of relevant risks that is not fully defeated by further facts of the case.¹¹

We should note that the cases we have considered so far comprise a special subset of cases in which an agent has very detailed beliefs concerning the likely consequences of her behavior. Dorfman believes that *Mrs. Dorfman will drink the arsenic-laced tea and die*. Carla believes that *Delores will be crushed when the house collapses*. Often, however, agents are blameworthy for consequences that they cannot articulate in such precise terms. Consider one more case.

Spleen: Stacy is speeding recklessly through a neighborhood and hits a pedestrian, Tanda, rupturing her spleen. Stacy has no idea that there is an organ called a spleen.

Stacy is blameworthy for hitting Tanda and rupturing Tanda's spleen. But unlike in the previous cases, it seems false to say that Stacy possesses evidence that she might rupture someone's spleen. Rather, Stacy knows that her act risks a broad *kind* of consequence, bodily injury to a pedestrian, and is blameworthy for Tanda's ruptured spleen because that outcome *realizes* the risk, in the sense that a ruptured spleen is an instance of bodily injury to a pedestrian that comes about in roughly the way Stacy expects (her car impacts a pedestrian, transfers momentum to the pedestrian's body, etc.).

¹¹ When is the support for a belief completely, rather than partially defeated? This is an interesting question that goes beyond the scope of the present discussion. In general, however, if an agent treats a certain fact as evidence on the basis of some further belief, the falsity of the supporting belief completely defeats the evidential quality of the fact in question. For instance, if one treats a certain calculation as evidence that a house will collapse, and does so because one takes the calculation to be correct, then the fact that the calculation is incorrect completely defeats the evidence one has that the house will collapse.

We can adopt this language to describe the two previous cases as well. Dorfman has evidence of a risk to Mrs. Dorfman's life (albeit a risk that the available evidence suggests is rather smaller than he believes it to be), a risk that is realized by her arsenic poisoning. Carla perhaps has some weak evidence that she risks causing Delores's death by not reinforcing the foundation of Delores's house (i.e., the fact that she came to this conclusion after doing her calculations), but the way Delores's death comes about does not realize any risk known to Carla, since any evidence Carla has is fully defeated by the further facts of the case. Together, these cases suggest the following necessary condition on blameworthiness for consequences.

Knowledge of Risk: An agent A is blameworthy for some consequence C of behavior X only if

- (i) A believes on the basis of evidence E that X might result in a consequence of kind K, and
- (ii) C realizes K in a way such that E is not completely defeated.

Knowledge of Risk offers a neat explanation of the cases we have seen so far. In Meeting for instance, Mary does not possess any evidence that arriving late to the cafe could result in Nina's death and so fails to meet condition (i); Knowledge of Risk thus correctly implies that Mary is blameless for Nina's death. In Concrete, even if we grant that Carla's mistaken calculation gives her some evidence that failing to reinforce the foundation could cause Delores's death such that condition (i) is satisfied, this evidence is completely defeated by the fact that her calculations involve a gross error, so that condition (ii) is not satisfied. Knowledge of Risk thus correctly implies that Carla is blameless for Delores's death. In Dorfman's case, Dorfman believes, on the basis of relevant evidence, that his actions might cause Mrs. Dorfman to die of arsenic poisoning, satisfying condition (i), and Mrs. Dorfman dies in precisely the way Dorfman's evidence suggests she might, satisfying condition (ii); Dorfman meets the conditions

on blameworthiness given by Knowledge of Risk. Finally, the same is true in Spleen. As a competent driver, (i) Stacy knows she risks harming a pedestrian, and (ii) Tanda's ruptured spleen is an instance of such harm.

I will discuss one more kind of case that gets particular attention in the philosophical literature. These are cases of *culpable ignorance*, cases in which an agent is ignorant that she is imposing a certain kind of risk but is blameworthy for said ignorance and so is blameworthy when the risk realized. Consider this case from Smith [1983].¹²

[A] doctor who treated a premature infant's respiratory distress in 1954 by exposing her to unnecessarily high concentrations of oxygen, and so caused severe eye damage, did something terrible. But if the doctor did not realize that oxygen enrichment would have this effect, he is not to blame for the baby's blindness ... However ... the latest issue of his medical journal described a study establishing this effect and recommending the use of lower concentrations as equally effective for respiratory problems. The doctor should have read this journal, and if he had done so, would have realized he ought to use less oxygen. (Smith [1983], 543)

The doctor in this case is plausibly blameworthy for blinding the infant. The standard view, suggested by Smith and others, is that although the doctor does not possess evidence that the treatment will blind the infant, the doctor is blameworthy for blinding the infant because he is blameworthy for failing to read the journal, which he should have done, and which would have provided him with information that would have enabled him to avoid this outcome. Such cases might seem to pose a problem for Knowledge of Risk since Knowledge of Risk implies that, if the doctor is blameworthy for blinding the infant, he must have possessed evidence of this risk.

¹² Holly Smith, "Culpable Ignorance", *The Philosophical Review*, 1983.

I believe that if one fully spells out why the doctor is culpable in this case, then the case in fact does not undermine, but provides support for Knowledge of Risk. The doctor is blameworthy for blinding the infant, on my view, because this specific consequence realizes a broader kind of risk that the doctor knows is imposed by his failure to read the journal: harm to one of his infant patients. The doctor's belief that a full strength dose of oxygen is appropriate, his use of too much oxygen, the infant's blindness — all of these states of affairs are blameworthy consequences of the doctor's earlier decision not to read the journal, at which time the doctor possessed evidence that his actions created the relevant kind of risk.¹³

To see that if the doctor is blameworthy for blinding the infant, he must have possessed evidence that he might injure an infant when he chose not to read the journal, consider a variation on the case.

Family: Felicity is a family doctor in a small town where she treats both elderly patients and young children. Given that she treats these patients, she is obligated to read two publications, *The Journal of Geriatric Medicine* and *The Journal of Pediatric Medicine*. One week, Felicity neglects to read *The Journal of Geriatric Medicine*, a decision for which she is seriously blameworthy.

Unfortunately, the issue of *Geriatric Medicine* that Felicity fails to read includes an insert that should have been included in *Pediatric Medicine*. (The journals are mailed together,

¹³ It is sometimes said that an action is *derivatively blameworthy*, blameworthy in virtue of being performed out of ignorance caused by a prior blameworthy act. For instance, in the case of the doctor, above, the act of giving too much oxygen is derivatively blameworthy in virtue of the prior blameworthy decision not to read the journal. On my view, this is the same as saying that the giving of too much oxygen is a blameworthy consequence of the earlier decision. (See Gideon Rosen, "Skepticism about Moral Responsibility", *Philosophical Perspectives*, 2004)

and a clerk placed the insert in the wrong one. No inserts have ever been included before, let alone included in the wrong journal.) The insert contains a recommendation that pediatric doctors lower the oxygen concentrations given to infants suffering respiratory problems, since the preliminary findings of a recent study suggest that there is a link between high oxygen concentrations and eye damage.

When Felicity encounters an infant with respiratory problems later in the week, she treats the infant with the traditional high oxygen concentrations, causing severe eye damage and blinding the infant. If Felicity had read *Geriatric Medicine*, she would have found the insert and avoided doing this damage.

Felicity is blameless for the infant's blindness. This is so, even though Felicity is blameworthy for failing to read *Geriatric Medicine*, just as Smith's doctor is blameworthy for failing to read the journal in the original case. The difference, on my view, is that in Felicity's case, the infant's blindness does not realize a kind of risk for which she had evidence when she failed to read the journal. When Felicity fails to read *Geriatric Medicine*, she does not possess evidence that she is risking harm to one of her pediatric patients.¹⁴ Thus, in addition to its other advantages, Knowledge of Risk also seems to underpin a correct account of when ignorance of a certain kind of risk is culpable or blameworthy. This counts strongly in favor of the view.

¹⁴ It is true that Felicity has evidence that she risks *Harming one of her geriatric patients*, and therefore has evidence that she risks *Harming one of her patients*, and further that *Harming one of her pediatric patients* is a way of harming one of her patients. However, because Felicity's evidence that she risks *Harming one of her patients* derives entirely from her evidence that she risks *Harming one of her geriatric patients*, the outcome of *Harming one of her pediatric patients* does not realize any risk known to Felicity.

1.2 General Epistemic Conditions

In the previous section, I argued that to be blameworthy for some consequence of her behavior, an agent must stand in a particular epistemic relationship to that consequence. This relationship is, I claim, given by Knowledge of Risk: an agent is blameworthy only for the consequences of her behavior that she believes might obtain on the basis of evidence that is not completely defeated by any further facts of the case.

On its own, however, Knowledge of Risk does not fully account for the role that an agent's epistemic state plays in explaining why she is blameworthy. In Meeting, for instance, the fact that Mary is blameworthy for the decision to continue watching her television show is explained by the fact that (i) doing so creates a risk that she will be late to her meeting with Nina, and (ii) Mary is aware of this risk. In Family, even though Felicity is blameless for the harm she does, she is nonetheless blameworthy for failing to read *Geriatric Medicine* because (i) the information in the journal is necessary for her to safely and effectively serve her patients, and (ii) she knows that the information is important in this way. The fact that an agent's epistemic state plays this explanatory role motivates a hypothesis that I call the General Explanatory Condition.

General Explanatory Condition: The fact that an agent A is blameworthy for behavior X is sufficiently explained by facts of the form E(A, P), where E is an epistemic attitude that A bears towards some morally significant proposition P.^{15,16}

¹⁵ Theorists who accept a control or freedom condition should perhaps formulate this hypothesis slightly differently.

General Explanatory Condition*: The fact that an agent A is blameworthy for behavior X is sufficiently explained by (i) the fact that A freely engaged in X, and (ii) facts of the form E(A, P), where E is an epistemic state that A bears towards some morally significant proposition P.

The argument that I offer against General Explanatory Condition below can be offered against General Explanatory Condition* *mutatis mutandis* by assuming that the agent in question acts freely.

For instance, some theorists maintain that what makes an agent blameworthy is that she engages in behavior that she knows to be morally wrong.¹⁷ Other theorists endorse a weaker condition on which an agent need only believe that her actions have certain features that, if they were to obtain, would make those actions wrong, regardless of whether the agent knows that these features are morally significant in this way.¹⁸ Both views hold that the agent's blameworthiness is sufficiently explained by her epistemic state, in the former case, the agent's knowledge that she acts wrongly, in the latter case, her morally-relevant beliefs.

Despite its relatively wide acceptance, I believe that the General Explanatory Condition is false. I claim that there are cases where (i) a certain fact plays an essential role in explaining why an agent is blameworthy, but (ii) this fact is not a constituent of the agent's epistemic state.

Consider the following case.

Interview: Irene is interviewing Janet for a position at a company. At one point, Janet demonstrates her qualification for the position by telling a story that, unbeknownst to Janet, vividly recalls for Irene some intensely embarrassing memories. As a result, although Irene is sensitive to Janet's strengths as a candidate, she leaves the interview with an unfavorable impression of Janet. Irene does not have the self-awareness to realize that her personal feelings might have an influence on how she evaluates Janet's candidacy.

¹⁶ There will also be an explanation for why the agent is blameworthy that involves, e.g., microphysical facts of which she is unaware. However, this does not prevent an explanation for why the agent is blameworthy given in terms of the agent's epistemic state from being a sufficient explanation.

¹⁷ An early version of this view is defended by Zimmerman [1997] and continues to figure in debates about the epistemic conditions on blameworthiness. (Michael J. Zimmerman, "Moral Responsibility and Ignorance", *Ethics*, 1997)

¹⁸ Rosen [2008] argues for such a condition, though he endorses a version of the stronger condition as well.

When Irene is later asked to evaluate Janet's suitability for the position as part of the hiring committee's group deliberation, Irene testifies that Janet may not perform well on the job and gives some inconclusive evidence to this effect. While Irene would not draw attention to this evidence unless it were relevant to the hiring group's deliberation, Irene is not motivated by a desire to hire the best candidate; her true motivation for calling attention to this evidence is simple dislike of Janet. However, Irene has no idea that this is what is motivating her to testify against Janet, nor could it occur to her.

I believe that Irene is blameworthy for giving testimony that Janet may not perform well on the job. This is not because of any fact about Janet's overt actions: In group deliberation, there is nothing wrong with calling attention to a subset of the available evidence and raising concerns supported by that evidence, even if that evidence is inconclusive.¹⁹ What makes Irene's action blameworthy is that it is motivated by a desire to malign Janet. That this is Irene's motivation is, however, opaque to her.

Cases like Interview show that General Explanatory Condition is false. The argument is as follows.

The No General Explanatory Condition Argument:

(P1) Irene is blameworthy for testifying against Janet.

¹⁹ This is not to say that Irene's behavior is permissible all things considered. I claim, first, that Irene's overt behavior need not be wrong if it were motivated out of concern to hire the best candidate, and second that, if it is wrong for Irene to testify as she does, her mental state is a necessary part of the explanation as to why.

- (P2) If General Explanatory Condition is true, then if Irene is blameworthy for testifying against Janet, the fact that Irene is blameworthy for testifying against Janet is sufficiently explained by the epistemic attitudes Irene has at the time.
- (P3) Irene does not bear any epistemic attitude towards the fact that she is motivated to testify against Janet out of dislike.
- (P4) Any sufficient explanation as to why Irene is blameworthy involves the fact that Irene is motivated to testify against Janet out of dislike.
- (C) General Explanatory Condition is false.

The No General Explanatory Condition Argument is valid. Are the premises true? I will take P1 for granted: Irene is blameworthy for testifying against Janet.²⁰ P2 is also true, as it is a tautology: the antecedent is the claim that General Explanatory Condition is true and the consequent is a substitution instance of General Explanatory Condition.

As a straightforward claim about the facts of the case, P3 is also uncontroversial. Irene is motivated to testify against Janet out of dislike but has “no idea” — no belief or credence, explicit or implicit, nor the ability to form any such belief — that this is her motivation. We should accept P3 unless it is implausible that an agent like Irene would be in this way unaware of her motivations. It is not implausible; agents do sometimes act on motivations without any attitude, conscious or unconscious, about their motives. Such reflection is simply not an essential part of human action.

P4 is the argument’s most controversial premise. The best evidence that Irene’s motivations for testifying against Janet are an essential part of the explanation as to why Irene is blameworthy is the way in which Irene’s blameworthiness evolves through the case. Imagine first that, after

²⁰ Some readers will disagree. I consider arguments that hold that Irene is blameless in section 1.3.

their initial meeting, Janet were informed that Irene came to dislike her because her story caused Irene to recall some intensely embarrassing memories. It would be reasonable for Janet to be upset about Irene's attitude and it would be reasonable for her to feel that Irene's attitude is unfair; after all, Irene's embarrassment isn't Janet's fault, since she had no way of knowing her story would embarrass Irene. It would be unreasonable, however, for Janet to resent Irene for having the negative emotional response to Janet that she has when she listens to Janet's story. Simply having such an emotional response is not the kind of thing for which one can reasonably blame someone.²¹

Now consider Irene's testimony against Janet. As indicated above, Irene's speech would not be blameworthy if, e.g., it had been given purely as part of an effort to determine whether Janet is the best candidate for the job. Imagine, however, that the fact that Irene was motivated out of simple malice were to come to light. In that case, it would be perfectly appropriate for Janet to feel resentful of Irene's malicious testimony and inappropriate if Irene did not feel seriously guilty. These emotions, guilt and resentment, are paradigmatic forms of blame, and they become appropriate when and because Irene is motivated by her dislike of Janet. This is good evidence that Irene is blameworthy for her testimony in virtue of her motivations. We should accept P4.

We should accept all of the premises of The No General Explanatory Condition Argument.

Apparently, then, the argument is sound. General Explanatory Condition is false.

²¹ While not crucial to the argument here, I think that the best explanation of this fact is that emotional reactions are involuntary, and that blaming someone involves a belief that the offending behavior was voluntary. My considered view is that involuntary psychological states — emotions, beliefs, desires — are only ever derivatively blameworthy, blameworthy in virtue of being the consequence of a blameworthy act or omission. Intentions and mental actions like the making of a decision, on the other hand, can be non-derivatively blameworthy. Chapter Two discusses the importance of voluntariness to moral blame.

Where does General Explanatory Condition go wrong? The idea that an agent's epistemic state plays an important role in determining whether she is blameworthy is surely correct. Even in Interview, Irene's knowledge that she is in an interview setting, that her testimony may affect the outcome of Janet's candidacy, etc., is part of the explanation for why her testimony is blameworthy; Irene's malicious testimony would have been much less blameworthy, and may not have been blameworthy at all, if she were speaking to a friend rather than to the hiring committee. In each of the cases we have considered, it seems that the blameworthy agent expresses a morally objectionable quality of will in part because she is aware of *moral reasons* that bear on her actions, facts that bear on whether an act is good or bad, right or wrong. This hypothesis, however, is not represented by General Explanatory Condition, but by a weaker condition.

Restricted Explanatory Condition: The fact that an agent A is blameworthy for behavior X is always partially explained by facts of the form E(A, M), where E is an epistemic attitude that A bears towards some moral reason M.²²

Whereas General Explanatory Condition implies that an agent has epistemic access to every fact that contributes essentially to her blameworthiness, Restricted Explanatory Condition only claims that she has epistemic access to some such facts. In particular, it allows for the possibility that the explanation as to why some agent is blameworthy involves both a moral reason to act otherwise of which she is aware and also some conflicting motive of which she is unaware.²³

²² The theory of blameworthiness advanced by Arpaly [2002] suggests that responsiveness or unresponsiveness to certain reasons always explains why an agent is blameworthy, even in cases like Interview. I endorse this view in the following chapter. (Nomy Arpaly, *Unprincipled Virtue*, Oxford University Press, 2002)

²³ This is not to suggest that one's motives never themselves constitute moral reasons to act one way or another, only that the moral reasons of which an agent is aware might not include her motives.

Irene's epistemic access to the moral reasons she has to contribute to the hiring committee's selection of the best candidate and to the stakes of her testimony for Janet's candidacy is plausibly essential to her being blameworthy. As argued in the previous section, epistemic access to the consequences of one's actions is essential to being blameworthy for those consequences. Epistemic access to one's own motivations, however, is not necessary. What explains this apparent asymmetry? The explanation, I believe, lies in the fact that blame indicts agents for behavior that expresses a *morally objectionable quality of will*, a lack of concern for the rights and interests of others.²⁴ When one blames an agent for some behavior, one is not merely criticizing that behavior, but also criticizing the agent's motivations for engaging in that behavior. An agent need not be aware of her motives in order for those motives to constitute a morally objectionable quality of will that is expressed in her behavior.

The idea that an agent's behavior can express a morally objectionable attitude is intuitive enough when we focus on a decision to act in one way or another: Millie's decision to continue watching television in Meeting, Dorfman's decision to poison Mrs. Dorfman, Carla's decision to seize the opportunity to kill Delores in Concrete; in each case the blameworthy agent expresses a lack of respect for someone else's rights or interests. However, we also blame agents for the consequences of their behavior, and it is harder to see how an outcome or consequence could express an agent's moral quality of will. Outcomes, after all, are just events or states of affairs. How can a state of affairs or an event express an agent's will?

The answer is just the view defended in the previous section: states of affairs and events express an agent's will when the agent is in a position to take their possibility into account in

²⁴ The idea that we blame agents for behavior that expresses a certain quality of will comes from Strawson [1962]. (P.F. Strawson, "Freedom and Resentment", *Proceedings of the Aristotelian Society*, 1962)

deciding how to act. The state of intending something, say, *Intending to serve someone a shiitake broth*, necessarily expresses an agent's will; it constitutes that will. But while the resulting action might, in virtue of its effects, satisfy the description *Causing someone to suffer a severe allergic reaction*, the complex action to which that description refers only expresses the agent's will if the agent is sensitive to evidence that her action risks initiating a sequence of events that would make that description of her action true.

The real problem with General Explanatory Condition, on my view, is that it elides two distinct ways that an agent's epistemic state is relevant to blameworthiness: On the one hand, an agent's epistemic state contributes to the moral quality of the will expressed in her behavior. On the other, her epistemic state determines which of the resulting behaviors express that will.

1.3 Objections

In this section, I consider some objections to the claims made in sections 1.1 and 1.2 that clarify the view and some of its further consequences.

Objection: Knowledge of Risk claims that an agent must have a belief that her actions impose a certain kind of risk in order to be blameworthy if such a risk is realized. But this is too restrictive; an agent might be blameworthy for an action and its consequences because, although she doesn't realize that her actions impose such a risk, she is capable of forming beliefs about the relevant risks and would have formed the relevant beliefs if she cared sufficiently about such risks.

Cases like the following might seem to support the objection.

Botulism: Brenda cans corn using a process shown to her by Courtney that involves cooking the sealed cans in a pressure cooker. Unfortunately, Brenda's pressure cooker isn't working, so she deliberates on whether putting the sealed cans in a bath of boiling water is an acceptable substitute process. Although Brenda could form the belief that boiling the cans in an open water bath likely poses a higher risk of foodborne illness than cooking them in a pressure cooker — and although she would form this belief if she cared appropriately about food safety — Brenda instead comes to the conclusion that boiling the cans in a water bath will reduce the risk of foodborne illness to the same acceptable levels as would a pressure cooker. She proceeds to boil her canned corn. Unbeknownst to Brenda, however, the boiling water bath is insufficient to kill the *botulinum* bacteria in her corn. When she later serves her daughter Delilah some of the canned corn, Delilah becomes seriously ill with botulism poisoning.

The objection relies on the intuition that an agent like Brenda is blameworthy for her decision to can corn using only a boiling water bath instead of waiting to use a pressure cooker, and is further blameworthy for causing Delilah to become seriously ill.

My view is that cases like this one are misleading. It is natural to think that Brenda doesn't really believe that the boiling water bath is just as safe as the pressure cooker or, even if she does, that she could nonetheless choose to refrain from proceeding because she doesn't have sufficient evidence to justify acting on this belief. However, if Brenda could refrain from her negligent behavior for this reason, then she must implicitly know that she doesn't have the evidence required to rule out an unacceptable risk of foodborne illness. If this is the case, then the intuition that Brenda is blameworthy poses no problem for the views defended in the previous two sections: Delilah's illness is an instance of a kind of consequence of which Brenda is aware and which constitutes a moral reason for her to refrain from acting as she does. The

fact that Brenda also believes that she will likely avoid such an outcome doesn't prevent Delilah's illness from realizing a kind of consequence that Brenda knows could result from her actions, however unlikely she thinks it is.²⁵

Because cases like Botulism can be misleading, the objection is best tested by cases where it is clear both that an agent could have come to the right conclusion about the risk of a possible outcome and also that she does not have any implicit knowledge of reasons to refrain from acting as she does. Consider this case.

Checkmate: Charmaine's sister, Daisy, has been kidnapped by Eve, and Eve will only release Daisy if Charmaine defeats Eve in a game of chess. Charmaine has never cared very much about Daisy, but she agrees to play because she knows she should. During the game, there comes a point where one move, move A, will cause Charmaine to eventually win and where another move, move B, will cause Charmaine to eventually lose. Charmaine could realize the need to make move A (she is a competent player and is broadly disposed to see such opportunities) but, although she is concentrating as best as she can, she comes to the mistaken conclusion that she should make move B. Charmaine honestly believes that all of the available evidence supports this conclusion. However, if Charmaine cared more about Daisy, she would have had the insight that she should make move A. As it is, she makes move B and loses the game, consigning Daisy to an indefinite term of imprisonment by Eve.

Whereas Brenda seems blameworthy for her decisions in Botulism, in Checkmate, it seems to me clear that Charmaine is blameless. The difference, I think, is that Checkmate makes clear

²⁵ This is to assume that Brenda's deliberation does not result in her being absolutely certain that boiling the cans eliminates any risk of foodborne illness. This assumption is uncontroversial: certainty is rare, especially in cases where an agent knows that her evidence is inconclusive.

that after deliberating to the conclusion that she should make move B, Charmaine has no idea, no implicit knowledge or belief, that her conclusion is unwarranted.

The reader might wonder how my view about cases like Checkmate is compatible with the conclusion that Irene is blameworthy in Interview. After all, both agents bring about bad outcomes because of some morally objectionable desire, in Irene's case, a desire to malign Janet, in Charmaine's, too weak a desire to rescue Daisy. The answer is that Charmaine's lack of concern for Daisy affects her decision making only indirectly; Charmaine's lack of concern is not a motivation for making move B, but a sub-personal causal factor that contributes to her failing to notice that she should make move A, which in turn affects her behavior. In Interview, on the other hand, Irene's dislike for Janet is her driving motivation for testifying against Janet.

My verdict in Checkmate has two implications. The first is that, contrary to the conclusion one might be inclined to draw from cases like Botulism, it is the recognition of a risk, not just the capacity to recognize a risk, that makes an agent blameworthy. The second is that whenever one blames an agent for inadvertently causing others harm, if the agent is not like Smith's doctor, not already blameworthy for her ignorance of the relevant risks, then either the agent must have some implicit appreciation of the relevant risks or one is making a mistake in blaming her.

Objection: It is pedantic to maintain that an agent who is motivated to act wrongly by a morally objectionable desire is blameworthy while an agent whose morally objectionable desires cause, but do not motivate her to act wrongly is blameless. Why should this fine distinction matter so much to the appropriateness of blame?

First, I disagree with the suggestion that the distinction between desires that cause an action and desires that motivate an action is overly scrupulous. It is, on my view, the only line to draw between cases where agents are blameworthy because of unconscious motives — whether those motives are one-off, like Irene’s, or systematic racist or sexist biases — and cases in which an agent is blameless because the causal role played by her morally objectionable desires is totally sub-personal, as is the case for Charmaine in Checkmate. Second, this distinction between desires that motivate objectionable behavior and desires that merely cause such behavior matters to blame, on my view, because blame is a natural reaction to an agent’s motives, not the sub-personal causes of her behavior.²⁶ Motives, and not brute causes, belong to the set of psychological concepts that it is natural to use in evaluating others’ behavior.

Objection: Because Knowledge of Risk allows for an agent to be blameworthy for very specific consequences in virtue of those consequences belonging to a very general kind, it implausibly allows agents to be blameworthy for consequences that are not “foreseeable” in any intuitive sense. This suggests that the epistemic conditions on blameworthiness in section 1.1 are too weak.

The following case might be thought to provide further support to the objection.

Reading: Regina receives a call from Samantha, who tells her that she left a note at Regina’s house and that Regina should go home and read it to prevent something bad from happening. Regina promises Samantha that she will read the note.

Regina has no idea what information is contained in the note or why Samantha made her promise to read it. In fact, the note contains information about how to prevent Tina’s

²⁶ I develop this view in detail in Chapter Two.

death. Regina fails to read the note and therefore does not take preventive action. Tina dies.

Intuitively, Regina is not blameworthy for Tina's death, but if this is so, Knowledge of Risk does not explain why. Regina possesses evidence that she risks something bad happening if she doesn't read the message, and Tina's death realizes this risk.

On my view, Regina is somewhat blameworthy for Tina's death, but much less blameworthy than she would be had she knowingly or intentionally brought about Tina's death. By this I mean that the forms of blame — and also the forms of punishment — that it is appropriate to direct at Regina are less intense than the extreme forms of condemnation and punishment that are normally appropriate when someone has, for instance, committed murder.²⁷ The same is true in cases like that of Smith's doctor, who creates a certain kind of risk by failing to read his medical journal but is ignorant of whether his actions will result in any serious harm: Smith's doctor is blameworthy for blinding the infant, but much less so than a doctor who, e.g., knew respiratory treatments for infants had improved but consciously chose not to investigate how he should change his practices. This doctor is in turn less blameworthy than someone who intentionally uses excessive oxygen to blind an infant. In each case, the intensity of blame and the severity of the punishment that is justifiable is significantly greater than the case before.

A similar scaling down of blameworthiness can be observed in cases where consequences of a certain kind come about in an unexpected way. Consider another case.

²⁷ Blameworthiness is widely believed to be a necessary condition for just punishment. Similarly, I think, the more intense the blame it is appropriate to feel towards someone, other things equal, the more severe the punishments that might be justified. The relationship between blame and punishment is discussed at length in Chapter Three.

Accident: Alexa is driving to a meeting and is running late. A light ahead of her turns yellow. She steps on the gas, runs the red, and hits a pedestrian, Beatrice, who has just stepped into the crosswalk. Beatrice, a surgeon, suffers a serious concussion and is unable to perform a lifesaving operation on Candice. Candice dies.

Alexa possesses evidence that her driving is dangerous to pedestrians. This makes her blameworthy for Beatrice's injuries. Further, one might think, injuries to pedestrians often negatively affect their friends, families, and associates by, for instance, subjecting them to trauma or financial stress. If Alexa would have been blameworthy for these consequences, why shouldn't she be blameworthy for Candice's death, which is, after all, just a particularly extreme instance of harm to a third party?²⁸

The best answer is that Alexa might be blameworthy for Candice's death, but is not blameworthy to anywhere near the degree usually implied by a phrase like "Blameworthy for So-and-so's death". How blameworthy is she? A good heuristic for assessing degrees of blameworthiness in such cases is to take the maximum degree of blameworthiness to be set by the paradigm instances of the kind of risk in question. In the case of harm to third parties in traffic accidents, these paradigms include the aforementioned emotional trauma, financial hardship, and so on. One can be seriously blameworthy for such consequences, but not blameworthy to the degree that usually accrues to someone guilty of manslaughter.

Objection: If degree of blameworthiness depends on both the quality of will expressed in an act and the badness of an outcome, Knowledge of Risk implausibly allows for blameworthiness to vary significantly in very similar cases.

²⁸ This is not to assume that Alexa would have been blameworthy for such consequences; it is an open question whether drivers typically stand in the right epistemic relationship to the possibility of harm to third parties to be blameworthy if such harms are realized.

This objection gets its force from pairs of similar cases. Consider Dorfman, who is blameworthy for poisoning Mrs. Dorfman. Dorfman is a murderer, and is thus *prima facie* seriously blameworthy. But in cases quite similar to Dorfman's that involve the same objectionable quality of will, but where the explanation for the death violates the conditions given by Knowledge of Risk, blameworthiness should then differ significantly. Consider a case.

Switch: Stephanie buys arsenic to use to poison Tamara. However, as Stephanie is preparing to poison Tamara, Uma switches Stephanie's arsenic for cyanide. (Uma wants Tamara dead and so substitutes a stronger poison.) Stephanie puts the cyanide in Tamara's tea. Tamara drinks the poisoned tea and dies.

Stephanie's evidence that she is poisoning Tamara — the fact that the stuff she is putting in Tamara's tea seems to be the arsenic she bought earlier — is completely defeated by Uma's substitution of the poisons. Thus, according to Knowledge of Risk, Stephanie is not blameworthy for Tamara's death, but rather is blameworthy only for attempting to poison Tamara. *Prima facie*, attempted murder is much less blameworthy than murder, and so Stephanie is much less blameworthy than Dorfman. The objection holds that this is implausible.

On my view, the similarity of the facts in Switch to the facts in Dorfman's case plays on a common dislike of the idea of moral luck. But any plausible theory of blameworthiness for consequences on which the degree of blameworthiness can increase with the badness of the consequences for which an agent is to blame will allow blameworthiness to depend on facts outside of an agent's epistemic possession. What is significant, from the perspective of Knowledge of Risk, is that Switch is structurally identical to Concrete. Both Stephanie and Carla succeed in killing someone, but as a kind of accident. Given that Dorfman is blameworthy for

Mrs. Dorfman's death and that Carla is blameless for Delores's, the fact that Knowledge of Risk treats Stephanie like Carla is a virtue of the view. If the reader is convinced that Stephanie and Dorfman are equally blameworthy, this is an objection to the claim that blameworthiness always increases with the badness of the consequences for which an agent is blameworthy, not an objection to Knowledge of Risk

Objection: Given that Irene has no idea that she is motivated to testify against Janet out of dislike and no ability to become conscious of her unconscious motives, then although Irene treats Janet unfairly, she is not blameworthy because she has no way to avoid acting as she does.

This objection raises a number of important issues. First is the question of whether, if Irene could not have avoided treating Janet unfairly, this fact renders her blameless. Whether incapacity is an excuse is an important issue that I discuss in Chapter Three. However, it is a mistake to infer from the fact that Irene has no access to what is motivating her the conclusion that she could not have acted any other way. When Irene is asked to testify, she plausibly retains the ability to, e.g., deliberate on whether Janet is the best candidate for the job, and to base her testimony on the results of that deliberation. Doing so would plausibly have prevented Irene from treating Janet unfairly out of dislike. For the sake of argument, let us assume that Irene could have so deliberated and that, if she had done so, she would have avoided treating Janet unfairly and hence would not have been blameworthy.

Does this answer the objection? One might worry that we now face a dilemma. On the one hand, Irene might be aware of moral reasons that require her to pause and deliberate. If this is the case, then Irene is plausibly blameworthy for failing to deliberate, and is blameworthy for her biased testimony because it is the consequence of that failure. In this case, Irene arguably

provides no counterexample to General Epistemic Condition since her blameworthiness is explained by facts of which she is aware: the facts that require her to deliberate. On the other hand, if Irene is not aware of reasons that require her to pause and deliberate, then perhaps she could not possibly have done so and is therefore blameless.

This is a false dilemma: Irene is plausibly aware of reasons that license, but do not require deliberation. The reasons that support trying to determine whether Janet is the best candidate — the reasons that should be motivating Irene — support deliberating on the question of whether Janet is the best candidate. But those reasons do not make deliberation mandatory: spontaneous testimony about Janet's faults would not be impermissible if it were motivated by a desire to determine, in group deliberation, whether those faults make Janet an unacceptable candidate.

What makes Irene blameworthy, on my view, is just that she fails to be motivated to take any of the numerous permissible courses of action available to her on the basis of reasons that justify so acting. Deliberation thus need not be obligatory in order to be a permissible option available to Irene that, if engaged in, would have prevented Irene from testifying against Janet out of dislike, and so prevented her from being blameworthy.

Objection: If Irene had the capacity to give testimony on the basis of the relevant moral reasons rather than out of dislike, then it is plausible that she is blameworthy not because she expresses a morally objectionable quality of will, but because she has this capacity and fails to exercise it. In general, the best explanation of why an agent's epistemic states matter to moral blame is not that they bear on whether she expresses a morally objectionable quality of will, but that they bear on whether she has the capacity to act permissibly.

This is an important objection, as it represents a fundamentally different approach to the issues of blame and responsibility than the one I take in this essay. This approach maintains not only that certain rational capacities are necessary for blameworthiness, but that the expression of a morally objectionable quality of will is not necessary for blameworthiness. To evaluate this claim, consider the following case.

Kitten: Kate is across the street from a burning building where firefighters are trying to rescue a trapped kitten. They do not know what room the kitten is in and the fire is spreading fast. Soon the firefighters will be forced to exit the building. Kate is extremely concerned for the kitten and wishes she could help.

Suddenly, Kate sees movement in a window and realizes that it's the kitten. Kate could call out to the firefighters and the kitten would be saved. All things considered, this is what Kate desires to do. However, Kate finds herself frozen with panic. A moment later, Kate loses her opportunity to act when the fire chief calls an exit from the building.

It is terrible that Kate fails to save the kitten. As stated in the case, Kate has the rational capacities necessary to take action rather than succumbing to panic. Her failure to do so, however, apparently involves no morally objectionable quality of will; all things considered, Kate wants to save the kitten.²⁹ If only the capacity to act rightly, not a morally objectionable quality of will, is required for blameworthiness, then Kate is blameworthy. But Kate is not blameworthy. As a would-be rescuer, she is a gross disappointment, but no one should resent her for her failure. The objection is false; if the rational capacity to act permissibly is a necessary condition for

²⁹ I discuss the relationship between an agent's quality of will and her desires further in Chapter Two.

moral blameworthiness, it is a condition that holds in addition to, not instead of a quality of will condition.

Objection: While the epistemic conditions argued for above can be grounded in concerns about the quality of will expressed in an agent's behavior, there are cases where (i) an agent's epistemic state explains why she is blameless for a certain outcome, but where (ii) this is not because her epistemic state prevents that outcome from expressing her will. Such cases suggest that, contrary to the story told in section 1.2, the epistemic condition on blameworthiness is not grounded solely in the significance that an agent's epistemic states have for the quality of will expressed in her behavior.

The objection is correct that I have not shown that an agent's epistemic state matters to blameworthiness only because of its relationship to quality of will. Other theorists argue, for instance, that the epistemic condition on blameworthiness is grounded in considerations of moral fairness. I have not argued directly against these views. However, by showing that a wide range of cases in which an agent's epistemic state apparently renders her blameless can be explained by how an agent's epistemic state contributes to the quality of will expressed in her behavior, I have presented a plausible alternative to such views.

There are at least two kinds of cases that might motivate an epistemic condition grounded in considerations of fairness rather than quality of will. On the one hand, there are cases of *moral ignorance*, cases where agents act wrongly on the basis of false beliefs about what is morally permissible. Consider this case from Rosen [2003].³⁰

³⁰ Gideon Rosen, "Culpability and Ignorance", *Proceedings of the Aristotelian Society*, 2003.

Ancient Slavery. In the ancient Near East in the Biblical period the legitimacy of chattel slavery was simply taken for granted. No one denied that it was bad to be a slave, just as it is bad to be sick or deformed. The evidence suggests, however, that until quite late in antiquity it never occurred to anyone to object to slavery on grounds of moral or religious principle. So consider an ordinary Hittite lord. He buys and sells human beings, forces labour without compensation, and separates families to suit his purposes. Needless to say, what he does is wrong. The landlord is not entitled to do these things. But of course he thinks he is. Moreover, we may imagine that if he had thought otherwise, he would have acted differently. (Rosen [2003], 64-65)

Some theorists claim that the Hittite lord's false moral beliefs render him blameless for his wrong acts, and render him blameless because it would be unfair to blame him. I disagree. The Hittite lord strikes me as clearly blameworthy for his actions, and blameworthy because of the terribly objectionable moral quality of will those actions express.³¹ If anything, such cases make me more confident that epistemic conditions on blameworthiness are grounded in the condition that an agent express a morally objectionable quality of will.

A different line of support for the objection comes from cases like the following.

Migraine: Millie is a doctor who has pioneered a kind of brain surgery that is, at present, the only way to alleviate a particularly severe kind of chronic migraine. Most patients are completely cured by the surgery, though a small number (about one percent) die as the result of the surgery even if the operation is performed perfectly. Because of the severe

³¹ Here I am in agreement with Harman [2011]. Most other commentators on the issue of moral ignorance, however, assume that whether blame is appropriate is a question of fairness. (Elizabeth Harman, "Does Moral Ignorance Exculpate?" *Ratio*, 2011)

nature of the migraines at issue, this risk is acceptable to the vast majority of patients with the relevant condition. Performing the surgery is, in general, morally permissible.

A relative of Millie's, Nadia, suffers these migraines and asks Millie to perform the surgery on her. Millie morally ought to perform the surgery. Millie, however, hates Nadia, and is happy to see her suffer. Millie would be even happier, however, to see Nadia dead. Millie therefore agrees to perform the surgery for Nadia, hoping that Nadia's is one of the rare cases in which the surgery results in death. Indeed, when Millie performs the surgery, Nadia dies on the operating table, despite Millie's having performed the surgery perfectly.

The objection runs as follows: (i) Millie is not blameworthy for Nadia's death, and her epistemic state — the fact that she is aware of reasons to perform the surgery that outweigh the small risk of death — explains why, but (ii) this is not because Millie's epistemic state prevents Nadia's death from expressing her will. If both of these claims are true, then it seems that an agent's epistemic state constrains blameworthiness in ways not explained by its relationship to quality of will.

Let's take these points in reverse order, beginning with the intuition that Nadia's death seems to express Millie's will. This is hard to deny. Millie wants Nadia dead and performs the surgery precisely because it makes Nadia's death more likely. Further, Knowledge of Risk allows for the possibility that Millie is blameworthy for Nadia's death, since Nadia's death comes about in precisely the way that Millie anticipates, for the reasons that she anticipates it will come about.

On the first point, however, it is not obvious that Millie is blameless for Nadia's death. What supports this intuition is a correlation between blame and punishment: other things equal, the

fact that one is blameworthy for some bad consequence indicates that one merits harsher punishments than if that consequence had not been realized. (This idea arguably underpins practices like punishing murder more harshly than attempted murder.) The reason that this correlation supports the objection is that Nadia's death does not seem to make Millie merit harsher punishment than she would have merited had Nadia survived: If before Nadia's death, Millie's objectionable motives merit only the punishment of, e.g., harshly-worded moral criticism, it seems preposterous that Nadia's death makes it appropriate to jail Millie as a murderer.³²

There are three ways one might respond to these points: (i) hold that Millie is not blameworthy for Nadia's death, the response assumed in the objection, (ii) hold that Migraine is a case where blameworthiness for consequences does not entail that an agent merits harsher punishment, or (iii) hold that it is not absurd that Nadia's death entails that Millie merits harsher punishments. Let us set aside (iii) and focus on (ii). Why not think that what this case shows is that there is not a perfect correlation between blameworthiness for consequences and the harshness of the sanctions an agent merits? Smith [1983] claims that the question of whether an agent is blameworthy for the consequences of her actions just amounts to the question of whether the realization of those consequences makes her more blameworthy, but this is far from obvious. Whether an agent is blameworthy for the consequences of an action can affect whether, e.g., one's blame should take the form of resentment or indignation, or whether one may — or must — forgive her. The fact that Millie is blameworthy not just for her bad intentions, but for Nadia's death might plausibly justify one continuing to blame her when forgiveness would otherwise be required.

³² It is an interesting question as to why this is. I discuss the relationship between blame and punishment more in Chapter Three.

These arguments do not settle what we should say about cases like Migraine, but they do show that the objection at least requires further support. Finally, note that even if the objection ultimately succeeds — as it might — it shows only that there is some other condition of blameworthiness to which an agent's epistemic states are relevant.³³ It does not show that the condition that an agent express a morally objectionable quality of will does not ground all of the epistemic conditions argued for above.

³³ My considered view of the fundamental conditions of blameworthiness, presented in Chapter Three, allows for this possibility.

2. Quality of Will

This chapter gives a precise account of when an agent expresses the kind of morally objectionable quality of will that is necessary for blameworthiness. I argue that such a quality of will is expressed when (i) an agent's beliefs and desires rationalize behavior that (ii) the agent possesses moral reasons not to engage in, reasons that (iii) the agent's motivations fail to counterbalance. The reason that blame is governed by these conditions, I argue, is that blaming someone involves having a thought about the offending agent's motivations that is true only if these conditions are met.

In “Freedom and Resentment”, P.F. Strawson suggests that the paradigm forms of moral blame — resentment, indignation, and guilt — are part of a set of *reactive attitudes*, attitudes that Strawson characterizes as follows.³⁴

The reactive attitudes I have so far discussed are essentially reactions to the quality of others' wills towards us, as manifested in their behaviour: to their good or ill will or indifference or lack of concern. (Strawson [1962], 56)

Attitudes of moral blame in particular are a natural response to behavior that seems to express what Strawson calls ill will or lack of concern, what I in this chapter call a *morally objectionable quality of will*.

³⁴ P.F. Strawson, “Freedom and Resentment”, *Proceedings of the Aristotelian Society*, 1962, as reprinted in *Perspectives on Moral Responsibility*, Cornell University Press, 1993.

Influenced by Strawson, many philosophers now take the expression of a morally objectionable quality of will to be a condition on whether moral blame is appropriate: an agent is only *blameworthy*, only an appropriate target of moral blame, for behavior that expresses a morally objectionable quality of will. One aim of this chapter is to make such claims more precise by answering two questions.

- (I) When does an agent's behavior express her will?
- (II) What makes the will expressed by an agent's behavior morally objectionable in the sense relevant to moral blame?

In the first section of this chapter, I argue that an agent's behavior expresses her will when it is rationalized by what she believes and desires. An agent's behavior expresses a morally objectionable quality of will when the agent possesses moral reasons to refrain from engaging in that behavior and is not motivated by considerations that counterbalance those reasons.

After answering these questions, however, there remains a question about why the expression of a morally objectionable quality of will — in the sense just given — is necessary for moral blame to be appropriate.

- (III) What about moral blame makes its appropriateness depend on an agent expressing a morally objectionable quality of will?

In the second section of the chapter, I argue that the answer to this question concerns the thoughts that are characteristically involved in moral blame. Part of what makes an attitude a form of blame, on my view, is that it involves a thought about the moral quality of the offending agent's motivations. This thought is what makes blame a "reaction", in Strawson's words, to

behavior that expresses a morally objectionable quality of will. I claim that such a characteristic thought is true — and blame therefore appropriate — only if the offending agent expresses a morally objectionable quality of will in the sense defined in the first section.

In the third section, I consider some objections to the views defended in the previous two sections. These include objections to the idea that the appropriateness of blame is explained by constituent thoughts, objections that hold that moral reasons for action cannot do the work the account requires of them, and objections to the analysis of what makes for the expression of a morally objectionable quality of will. My aim in responding to these objections is to further clarify the view and to illustrate some of its further consequences.

2.1 Moral Reasons and Rational Control

An agent's will is only expressed in behavior over which she has an appropriate kind of rational control. Consider a case.

Kitten: Kate is across the street from a burning building where firefighters are trying to rescue a trapped kitten. They do not know what room the kitten is in and the fire is spreading fast. Soon the firefighters will be forced to exit the building. Kate is extremely concerned for the kitten and wishes she could help.

Suddenly, Kate sees movement in a window and realizes that it's the kitten. Kate could call out to the firefighters and the kitten would be saved. All things considered, this is

what Kate desires to do. However, Kate finds herself frozen with panic. A moment later, Kate loses her opportunity to act when the fire chief calls an exit from the building.³⁵

It is terrible that Kate fails to alert the firefighters in time to save the kitten. However, although Kate could and should call out, her failure to do so does not express a morally objectionable quality of will. As a kind of involuntary glitch, it does not express her will at all. Thus, while Kate's inaction may make her a terrible disappointment, no one should resent her.

It is a salient feature of the case that Kate does not fail to call out by choice. Her inaction is a pure *omission*, as opposed to behavior performed on the basis of an *intention* to act in one way or another. Cases like Kate's can thus encourage the view that only intentional behavior expresses an agent's will. But this is false. Consider another case.

Security: Sarah is a security guard at a hotel. One night Sarah is standing at her post in the hotel lobby when she witnesses a mugging take place across the street. Because Sarah has no desire that would be satisfied by helping the victim of the mugging, she continues to stand at her post, not at all motivated to act. Sarah does not choose not to act; her inaction is a pure omission.

Sarah's failure to intervene in the mugging in some way — to sound an alarm, call the police, or perform any of the acts that could be reasonably expected to help the victim of the mugging — expresses a morally objectionable quality of will and is seriously blameworthy. This is true

³⁵ Kate "could" call out to the firefighters in the sense that she has the rational capacity to do so. As Smith [2003] points out, this usage of "could" is quite common; for instance, if one has the rational capacity to come up with the answer to some question but fails to do so, one might say, "I could have come up with the answer." I discuss rational capacities and their significance to moral blame more later in this section and in Chapter Three. (Michael Smith, "Rational Capacities or: How to Distinguish Recklessness, Weakness, and Compulsion", *Weakness of Will and Practical Irrationality*, Oxford University Press, 2003)

despite the fact that, like Kate's inaction in Kitten, Sarah's inaction is a pure omission. This case thus shows that an agent's behavior need not be the product of an intention in order to express her will.

Whether an agent's behavior expresses her will, I maintain, depends on whether it is *rationalized* by her preferences, that is, whether the behavior is permitted by the agent's capacity for means-end rationality, given what she believes and desires. Almost all intentional actions are rationalized by an agent's preferences; behavior that is not rationalized in this way — behavior like a reflex, for instance, or a tic — is generally recognized as involuntary. On my view, it is precisely because tics and reflexes are not products of this capacity for means-end rationality that agents are generally blameless if a tic or reflex results in insult or injury.³⁶

With respect to Kate's inaction in Kitten, given what we know about Kate's preferences, we know that her beliefs and desires do not rationalize her behavior; she would prefer to call out and save the kitten. This, I claim, is the reason that we recognize her inaction as a kind of glitch, one that does not express her will. On the other hand, Sarah's failure to intervene in the mugging in Security does conform to the standards of means-end rationality: the fact that Sarah has no desire that would be satisfied by helping the victim means that her inaction is not means-end irrational.

It is important to note that an agent's preferences might rationalize her engaging in some behavior that has features of which she is not aware. In such circumstances, only a subset of

³⁶ Tics and reflexes sometimes are the product of an agent's beliefs and desires. For instance, when one flinches from a blow, this reflex is explained by a preference to avoid being struck. An important fact about reflexes, however, is that they are largely insensitive to other preferences; even if one's life depends on not flinching, one might struggle to avoid doing so. The capacity for means-end rationality that rationalizes acts and omissions, on the other hand, is broadly sensitive to all of an agent's beliefs and desires.

the behavior in which the agent has engaged expresses her will. For instance, recall a case from Chapter One.

Shiitake: Suzy is making a shiitake mushroom broth to have for dinner with Tiffany.

Unbeknownst to Suzy, Tiffany is dangerously allergic to shiitake. Suzy has never heard of a shiitake allergy, nor does Tiffany inquire into the provenance of the broth before beginning to eat. As a result, Tiffany suffers a severe allergic reaction.

Suzy has no belief, implicit or explicit, that serving Tiffany shiitake broth risks causing her to suffer an allergic reaction. For this reason, Suzy's preferences do not rationalize *Causing Tiffany to have an allergic reaction*, though they do rationalize *Serving Tiffany shiitake broth*.³⁷

It is important that an agent's preferences can rationalize behavior that is, all things considered, not rational. Consider a case.

Weakness: Wanda is in a footrace against Xiang and loses by a hair. Xiang, however, has her eyes closed as she crosses the finish line. Xiang asks Wanda, "Who won?" If Wanda were to focus on the reasons to answer honestly — her desire to be an honest person, the risks of lying, etc. — she would tell the truth. Further, Wanda knows this about herself; she knows that she rationally ought to tell the truth and that she will regret lying. However, as Wanda prepares to answer Xiang, Wanda is very strongly tempted to claim victory for herself. "I won," Wanda lies.

³⁷ I am neutral on the metaphysical question of whether *Serving Tiffany shiitake broth* and *Causing Tiffany to have an allergic reaction* are two descriptions of the same act or whether the latter description includes events (e.g., Tiffany's becoming short of breath, etc.) not included in the first.

Relative to all of the reasons that bear on how Wanda should answer Xiang, including facts about the various desires she would form if she deliberated properly, the choice to lie is less than ideally rational.³⁸ Wanda's decision to lie, however, is nonetheless rationalized by the preferences that she actually has while she is answering Xiang, a moment where her desire to claim victory is very strong and her competing desires are relatively weak. Thus, while Wanda's values may support acting rightly, on this occasion she expresses a morally objectionable quality of will.³⁹

Let us turn to the question of what makes the quality of will an agent expresses morally objectionable. Consider Wanda's lie in Weakness. What does it mean to say that Wanda's will here is morally objectionable? One natural thought is that her lie expresses a morally objectionable quality of will because it is wrong to lie in these circumstances. It is morally impermissible to lie for no good reason, and Wanda has no good reason for lying, no greater good that she is trying to serve, no secret she is obligated to protect, etc.

This answer, however, seems incomplete when one considers cases in which agents are blameworthy for behavior that is in some "objective" sense morally permissible. Recall a case from the previous chapter.

³⁸ What is involved in deliberating properly? This is a hard question. It seems that part of the explanation for why Wanda is not deliberating properly is that her desires to be an honest person, avoid the risks of lying, etc. are, at the moment, quite weak, but would become stronger if she were to focus her attention on such considerations. Lewis [1989] emphasizes the importance that imaginative acquaintance with relevant considerations has for deliberation. Lewis's theory, however, grounds an agent's reasons in her second-order desires, a view that founders on cases like Weakness: the reason for Wanda to focus her attention on the reasons she has to tell the truth is not grounded in a second-order desire to form such desires, since Wanda apparently has no such desire, at least not one strong enough to rationalize taking the required mental actions. Some alternative explanation is required, but this is beyond the scope of this essay. (David Lewis, "Dispositional Theories of Value", *Proceedings of the Aristotelian Society*, supplementary volume 63, 1989)

³⁹ This case illustrates how, although blame is an evaluation of an agent's will, it is distinct from characterological evaluation. Wanda might have a moral character that is quite excellent, but this is not directly relevant to whether she is blameworthy for lying to Xiang on this particular occasion. See Knobe [2002] for a discussion of related issues. (Joshua Knobe, *Lack of Character: Personality and Moral Behavior*, Cambridge University Press, 2002)

Interview: Irene is interviewing Janet for a position at a company. At one point, Janet demonstrates her qualification for the position by telling a story that, unbeknownst to Janet, vividly recalls for Irene some intensely embarrassing memories. As a result, although Irene is sensitive to Janet's strengths as a candidate, she leaves the interview with an unfavorable impression of Janet. Irene does not have the self-awareness to realize that her personal feelings might have an influence on how she evaluates Janet's candidacy.

When Irene is later asked to evaluate Janet's suitability for the position as part of the hiring committee's group deliberation, Irene testifies that Janet may not perform well on the job and gives some inconclusive evidence to this effect. While Irene would not draw attention to this evidence unless it were relevant to the hiring group's deliberation, Irene is not motivated by a desire to hire the best candidate; her true motivation for calling attention to this evidence is simple dislike of Janet. However, Irene has no idea that this is what is motivating her to testify against Janet, nor could it occur to her.

Irene expresses a morally objectionable quality of will in testifying against Janet. This is true despite the fact that it is in general permissible to raise concerns about a candidate in the kind of group hiring deliberation described in the case, even if those concerns are inconclusive. Irene's testimony does not express a morally objectionable quality of will because of its content; rather, what makes the quality of will expressed in Irene's testimony morally objectionable is her motivation for testifying against Janet, that is, the fact that Irene is motivated to testify against Janet out of dislike.

One could argue that Irene's motivations make her action morally impermissible, that it is wrong to testify against a job candidate out of dislike, endorsing something like the following principle.

Wrong-Making Motivations: If an agent A performs an act X with motivations M such that M makes X morally wrong, then A's performance of X expresses a morally objectionable quality of will.

Even if true, however, this claim raises the question of when an agent's motivations make it wrong to act as she does. One could insist that an account of the quality of will condition on moral blameworthiness need not include an answer this question, that the issue of what motives make an action wrong belongs to a separate theory of moral permissibility. However, this response is unsatisfying. A theory of the kind of morally objectionable quality of will required for blameworthiness needs, if not a complete theory, at least a schematic account of the kinds of motivations that make for the expression of a morally objectionable quality of will.

One approach to providing such an account is suggested by the work that Arpaly [2002] and Markovits [2010] have done on the Kantian concept of "moral worth", a cousin to Strawson's notion of quality of will.⁴⁰ Both Arpaly and Markovits take the moral worth of an action to be constituted by a relationship between the *moral reasons* relevant to the action — considerations that bear on whether the act is right or wrong, good or bad — and the agent's motivations. For instance, consider Arpaly's influential discussion of Mark Twain's Huckleberry Finn, who allows a black man, Jim, to escape from slavery in the American South.

⁴⁰ Nomy Arpaly, *Unprincipled Virtue*, Oxford University Press, 2002.
Julia Markovits, "Acting for the Right Reasons", *The Philosophical Review*, 2010.

At a key point in the story, Huckleberry's best judgment tells him that he should not help Jim escape slavery but rather turn him in at the first available opportunity. Yet when a golden opportunity comes to turn Jim in, Huckleberry discovers that he just cannot do it and fails to do what he takes to be his duty, deciding as a result that, what with morality being so hard, he will just remain a bad boy (he does not, therefore, reform his views: at the time of his narrative, he still believes that the moral thing to do would have been to turn Jim in). (Arpaly [2002], 9)

Arpaly argues that Huck is praiseworthy for allowing Jim to escape because this action has positive moral worth, because, even though he believes it is wrong to let Jim escape, Huck is motivated by a recognition of Jim's moral status, the facts about Jim that make it right to help him escape slavery. Conversely, blameworthy agents fail to be motivated by the moral reasons that justify their actions, whether out of indifference to the moral good or a desire for something morally bad.

In the spirit of such views, I offer the following account of when an agent expresses a morally objectionable quality of will.

Countervailing Reasons: An agent A expresses a morally objectionable quality of will in behavior X if and only if

- (i) A's preferences rationalize Xing,
- (ii) A possesses moral reasons M to refrain from Xing, and
- (iii) A's motives for Xing do not counterbalance M.

I will first review the evidence that each of these three conditions is necessary for the expression of a morally objectionable quality of will, then review the evidence that suggests that they are jointly sufficient.

Condition (i) is true since an agent's behavior can only express a morally objectionable quality of will if it expresses her will in the first place. Condition (i) explains why Kate's inaction in Kitten does not express a morally objectionable quality of will.

Condition (ii) captures the idea that morally objectionable behavior is behavior that an agent has some moral reason not to perform. To possess a moral reason that bears on some particular behavior, it is generally necessary for one's means-end reasoning for that behavior to be sensitive to the relevant facts.⁴¹ Thus in Shiitake, although the fact that Tiffany is allergic to shiitake is a reason for Suzy to refrain from serving Tiffany shiitake broth, Suzy is blameless for serving the broth because she does not possess this reason (or any other reason to refrain from doing so). Condition (ii) explains why Suzy does not express any morally objectionable quality of will in serving Tiffany the broth.

Condition (iii) is necessary to account for cases in which an agent possesses moral reasons to refrain from some behavior but is nonetheless justified in engaging in this behavior. Consider a case.

Factory: Francine is an executive at a materials manufacturer. She has just received a report that if she lays off scores of workers at a factory, she can afford to invest in new, automated technology that will reduce the factory's environmental impact by 99%. The

⁴¹ I will also assume that an agent possesses a moral reason if she is blameworthy for failing to countenance that reason. See my discussion of cases of "culpable ignorance" in Chapter One, section 1.1.

environmental benefits outweigh the interests of the workers. Although Francine wishes she could retain the workers, she lays them off so that she can invest in the new technology and protect the environment.

Although Francine possesses a moral reason not to lay off the workers — the hardship this will cause them — Francine is not blameworthy for doing so. On my view this is because, motivated as she is by the greater good of the environment, Francine does not express a morally objectionable quality of will.

Countervailing Reasons predicts that Kate from Kitten, Suzy from Shiitake, and Francine from Factory do not express a morally objectionable quality of will. Each fails to meet one of the three conditions given above, supporting the claim that each of these conditions is in fact necessary.

Interview provides an important test of the sufficiency of Countervailing Reasons, since Irene intuitively expresses a morally objectionable quality of will despite the fact that overt testimony of this kind is not necessarily wrong. I claim that, like most job applicants, Janet's interest in being hired constitutes a moral reason to encourage the hiring committee to hire her, not to discourage them. Of course, such moral reasons are often outweighed by other reasons, e.g., to only hire someone who will contribute adequately to the company, to hire the best candidate for the job, etc. If Irene were motivated by such considerations, her situation would be like Francine in Factory and she would not express a morally objectionable quality of will. However, Irene is motivated by simple dislike of Janet. Thus, since (i) Irene's preferences rationalize testifying against Janet, (ii) Irene possesses moral reasons not to testify against Janet, and (iii) Irene is not motivated by considerations that counterbalance these moral reasons, she expresses a morally objectionable quality of will.

It is worth reviewing the results that Countervailing Reasons gives in a few more cases.

Consider a variation on the Factory case.

Factory, Take Two: Fiona is an executive at a materials manufacturer. She has just received a report that if she lays off scores of workers at a factory, she can afford to invest in new, automated technology that will reduce the factory's environmental impact by 99%. The environmental benefits outweigh the interests of the workers. Because Fiona falsely believes that she has no moral duty to protect the workers' jobs it is easy for her to decide to invest in the new technology and protect the environment.

Countervailing Reasons implies that Fiona does not express a morally objectionable quality of will in laying off the workers — at least, not the kind necessary for moral blame to be appropriate — since, although the interests of the workers give her moral reason to keep them on, Fiona's reasons for laying off the workers (the environmental benefits) ultimately do outweigh the cost to the workers.

One might object that the fact that Fiona sees no reason to help her employees does seem to express a morally objectionable quality of will. Is this a counterexample to Countervailing Reasons? I maintain that it is not. Countervailing Reasons plausibly explains why it would be unreasonable for the factory workers to blame Fiona for her decision: shutting down the factory is supported by the environmental benefits. If the factory workers blame Fiona for anything, they should blame her for dismissing their interests, something that, contrary to Fiona's false moral beliefs, she should not do. Countervailing Reasons can account for this verdict given that (i) Fiona's beliefs and desires rationalize dismissing the workers' interests, (ii) Fiona possesses

moral reasons to take the workers interests seriously, and (iii) Fiona has no good reason for dismissing their interests.⁴²

Consider one more version of the case.

Factory, Take Three: Florence is an executive at a materials manufacturer. She has just received a report that if she lays off scores of workers at a factory, she can afford to invest in new, automated technology that will reduce the factory's environmental impact by 99%. The environmental benefits outweigh the interests of the workers. Florence has no interest in reducing her environmental impact, but she has been looking for a reason to punish the workers, who belong to a union that she detests. Florence announces that she is laying off the workers at the factory in order to protect the environment, but everyone knows that her real motivation is to punish the workers.

In this version of the case, it seems to me that it would be appropriate for the workers to resent Florence for the layoffs.⁴³ Countervailing Reasons explains why, since (i) the decision to lay off the workers is rationalized by Florence's beliefs and desires, (ii) the workers interests constitute moral reasons for Florence to not lay off the workers, and (iii) Florence is not motivated by considerations that counterbalance these reasons; in fact, Florence is positively motivated to bring about this bad result, expressing a kind of ill will for which resentment seems particularly appropriate.⁴⁴

⁴² If dismissing the factory workers' interests was not rationalized by Fiona's preferences because, say, she formed this belief spontaneously, then her false moral beliefs are bad but not blameworthy. Not all moral criticism amounts to blame, a fact I discuss further in Chapter 3, section 3.3.

⁴³ It is a further question whether the workers should blame Florence for the consequences of this decision, their job losses or decreased income, since they would have suffered these consequences even if Florence had acted as she should. See Chapter One section 1.3 for a discussion of this issue.

⁴⁴ Arpaly [2002] maintains that, other things equal, agents are more blameworthy for behavior motivated by ill will than behavior that is the result of mere indifference or lack of concern.

Finally, it is worth considering what Countervailing Reasons implies about cases in which an agent's behavior is the product of *compulsive* desires, desires that an agent is incapable of resisting. There are two cases worth considering.

Addiction One: Alice is supposed to pick her daughter Brittany up from school when she is struck with a craving for a narcotic a doctor prescribed her after a recent surgery. Alice knows that if she takes the narcotic, she will be unable to pick Brittany up from school, leaving Brittany stranded for hours. However, Alice's addictive desire is so powerful that she is incapable of resisting. Although she wants to care for Brittany more than she wants the pleasure of the drug, Alice's addiction drives her to take the drug.

It is hard to imagine the mindset described in this case: having desires best served by one course of action (picking Brittany up from school) rather than another (taking the narcotic), having all of the relevant information at one's disposal, but nonetheless taking the less rational course of action. It is a subject of some debate whether such cases are possible. However, if they are, if an agent can be compelled to act against her own strongest desires, then these are cases where agents are instrumentally irrational in the sense that their preferences do not rationalize behaving as they do. According to Countervailing Reasons, then, Alice's behavior is like a tic: it is a kind of psychological glitch that does not express her will.

Consider a second case.

Addiction Two: Anne is supposed to pick her daughter Belle up from school when she is struck with a craving for a narcotic a doctor prescribed her after a recent surgery. Anne knows that if she takes the narcotic, she will be unable to pick Belle up from

school, leaving Belle stranded for hours. However, Anne's addictive desire for the drug is so powerful that she is incapable of resisting. Although she knows it is wrong to leave Belle at school, her powerful desire for the narcotic outweighs her desire to pick Belle up from school, rationalizing the decision to take the drug.

In contrast to the verdict in the previous case, Countervailing Reasons implies that Anne expresses a morally objectionable quality of will, since, (i) per the stipulations in the case, Anne's beliefs and desires rationalize her taking the drug, (ii) her knowledge that she is leaving Belle stranded gives her strong moral reasons to refrain from taking the drug, and (iii) her obligation to Belle is not counterbalanced by her reasons for taking the drug. Simply put, it is objectionable that Anne's commitment to her daughter is not stronger than her desire to take the narcotic, and her actions express this objectionable preference. The fact that she could not resist her addictive desires if she wanted to is irrelevant.

One might think that even in this case blame is too harsh. After all, Anne's unfortunate desire to take the drug is the result of having had a doctor prescribe her a narcotic with powerful addictive properties. One might think that such circumstances are exculpatory. However, it is important to distinguish three questions: first, the question of whether an agent's actions express a morally objectionable quality of will; second, the question of whether an agent is blameworthy; and third, the question of whether it would be fair of us to blame the agent, all things considered.

Countervailing Reasons only implies an affirmative answer to the first question; it does not imply an affirmative answer to the second because Countervailing Reasons gives only a necessary, not a sufficient condition for blameworthiness. And countervailing reasons certainly does not imply an affirmative answer to the third question, since whether it is right to blame depends both on whether an agent is blameworthy and on various other factors. Thus, while Countervailing

Reasons implies that Anne's behavior expresses a morally objectionable quality of will, it leaves open the broader questions about what attitude we should take towards Anne.

Regardless of one's view on whether Anne is blameworthy, one should allow that she expresses a morally objectionable quality of will in choosing to take the narcotic. Consider one more case, a real life example introduced to the philosophical literature by Watson [1987].⁴⁵

On July 5, 1978, John Mayeski and Michael Baker had just driven through [a] fast-food restaurant and were sitting in the parking lot eating lunch. [...] At the other end of the parking lot, Robert Harris, 25, and his brother Daniel, 18, were trying to hotwire a [car] when they spotted the two boys. The Harris brothers were planning to rob a bank that afternoon and did not want to use their own car. When Robert Harris could not start the car, he pointed to the [car] where the 16-year-olds were eating and said to Daniel, "We'll take this one."

He pointed a ... [pistol] at Mayeski, crawled into the back seat, and told him to drive east...

Daniel Harris followed in the Harrises' car. When they reached a canyon area ... , Robert Harris told the youths he was going to use their car in a bank robbery and assured them they would not be hurt. [...] As the two boys walked away, Harris slowly raised the [pistol] and shot Mayeski in the back. [...] Harris chased Baker down a hill into a little valley and shot him four times. (Watson [1987], 235-236)

⁴⁵ Gary Watson, "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme", *Responsibility, Character, and the Emotions*, Cambridge University Press, 1987, as reprinted in *Agency and Answerability: Selected Essays*, Oxford University Press, 2004.

Harris's cruelty is explained by his history.

Harris was born Jan. 15, 1953, several hours after his mother was kicked in the stomach. She was 6½ months pregnant and her husband, an insanely jealous man, ... came home drunk and accused her of infidelity. [...]

Harris had a learning disability and a speech problem, but there was no money for therapy. When he was at school he felt stupid and classmates teased him, his sister said, and when he was at home he was abused.

"He was the most beautiful of all my mother's children; he was an angel," she said. "He would just break your heart. He wanted love so bad he would beg for any kind of physical contact. [...]"

[At age 14] Harris was sentenced to a federal youth detention center [for car theft]. He was one of the youngest inmates there, Barbara Harris said, and he grew up "hard and fast."

... Harris was raped several times, his sister said, and he slashed his wrists twice in suicide attempts. He spent more than four years behind bars as a result of an escape, an attempted escape, and a parole violation. [...] By the time he was released from federal prison at 19, all his problems were accentuated. Everyone in the family knew that he needed psychiatric help.

The child who had cried at the movies when Bambi's mother dies had evolved into a man who was arrested several times for abusing animals. (*Ibid.*, 239-241)

It is against this background that Harris murders Mayeski and Baker.

What attitude should we take towards Harris? He is deeply pitiable, a person who has suffered terribly and whose rage at the world is, if not justified, certainly understandable. At the same time, we should not deny that his murder of Mayeski and Baker expresses a terribly objectionable attitude towards the value of human life, a terribly objectionable moral quality of will. By distinguishing the three questions above (whether an agent expresses a morally objectionable quality of will, whether an agent is blameworthy, and whether we ought to blame her) we can acknowledge that it is clear that Harris's actions express a morally objectionable quality of will and at the same time remain open to different answers concerning the harder questions about whether Harris is blameworthy and the attitude that, ultimately, we should take towards him. My suggestion is that the same is true of Anne in Addiction. Her desire for the narcotic is stronger than her desire to fulfill her obligations to her daughter, and this is morally objectionable. Whether Anne is blameworthy and whether we should blame her are, however, further questions.⁴⁶

2.2 Quality of Will and the Nature of Blame

In the previous section, I gave an account of when some behavior expresses a quality of will that is morally objectionable in the sense required for moral blame to be appropriate.

⁴⁶ My view, discussed in Chapter Three, is that the question of whether Anne and Harris are blameworthy depends on whether there are moral principles of a certain kind that prohibit punishing them for their behavior. Because I am skeptical that there is any principle that prohibits punishing agents for acting on objectionable desires that are acquired as the result of bad luck — as Anne and Harris's certainly are — my own view is that Anne and Harris are blameworthy for their actions.

Countervailing Reasons: An agent A expresses a morally objectionable quality of will in behavior X if and only if

- (i) A's preferences rationalize Xing,
- (ii) A possesses moral reasons M to refrain from Xing, and
- (iii) A's motives for Xing do not counterbalance M.

Even if our discussion in the previous section has convinced the reader that expressing a morally objectionable quality of will, in the sense given by Countervailing Reasons, does in fact constitute a necessary condition on blameworthiness, one might reasonably wonder why this condition must be met in order for an agent to be blameworthy. This brings us to the third question from the introduction and what is perhaps Strawson's greatest contribution to the literature on moral blame, the thesis that blame is a reaction to the quality of will expressed in an agent's behavior.

According to Strawson, the question of whether an agent is blameworthy for some act must be answered in the context of the various attitudes that constitute moral blame. These attitudes paradigmatically take the form of resentment in cases where one is the victim of some moral wrong, indignation in cases where one witnesses a moral wrong, and guilt in cases where one has oneself perpetrated a wrong. If guilt, resentment, and indignation all are forms of blame, Strawson suggests, then a claim about the nature of blame is a claim about features common to these emotions.

Like all emotions, guilt, resentment, and indignation are complex psychological states that involve cognitive elements (thoughts and beliefs), conative elements (desires or other pro-attitudes), and affect (feelings and phenomenology). The constituent that would most obviously

account for Strawson's observation that blame is a reaction to expressed quality of will is a cognitive constituent, a thought about the quality of will expressed in an agent's behavior.

I offer the following conjecture about the nature of blame.

- Characteristic Thought:** If an attitude E directed at an agent A for behavior X is a way of blaming A for X, then E involves a cognitive element the content of which is true only if
- (i) A's preferences rationalize Xing,
 - (ii) A possesses moral reasons M to refrain from Xing, and
 - (iii) A's motives for Xing do not counterbalance M.

Characteristic Thought postulates that the attitudes that constitute forms of blame have a certain essential feature, namely, cognitive elements that imply the conditions that are, according to Countervailing Reasons, sufficient for the expression of a morally objectionable quality of will.

On my view, the fact that the expression of a morally objectionable quality of will in the sense given by Countervailing Reasons is a condition on moral blameworthiness is explained by the following principle.

- Cognitive Conditions:** If an attitude A involves a cognitive element with content P, then A is appropriate only if P is true.

Principles like Cognitive Conditions are widely endorsed.⁴⁷ They are endorsed in part because they have significant explanatory power. Why, for instance, is it inappropriate to be afraid when

⁴⁷ Hieronymi [2004], Graham [2012], Rosen [2015] and others endorse versions of this principle. (Pamela Hieronymi, "The Force and Fairness of Blame", *Philosophical Perspectives*, 2004; Peter

one is not in danger? Because fear is an attitude that involves thinking that one is in danger, and attitudes that involve false thoughts are inappropriate.

It is of course possible to fear a spider while knowing that the spider is not dangerous — or, for that matter, to blame someone for her behavior while believing that she was unaware of any moral reason to act otherwise. Do such cases conflict with the claim that fear of a spider involves a thought that the spider is dangerous or that blame involves the thought that an agent possessed moral reasons to act otherwise? No. Following Rosen [2015], I use the term “thought” to pick out a cognitive state similar to belief in that it presents the world as being a certain way, but that can coexist with conflicting or contradictory thoughts. It is possible for one to both have the thought some act expressed a morally objectionable quality of will while at the same time believing that this thought is mistaken, even though it may not be possible to simultaneously believe that an act has expressed an objectionable quality of will while at the same believing the opposite. Similarly, one can have a thought with the content *That spider is dangerous* despite knowing that thought to be false. Emotions are not the only source of cognitive states with conflicting content: in the case of perception, for instance, one can view the Müller-Lyer illusion and perceive one line to be shorter than the other even though one knows the lines to be of the same length (the fact that one line appears longer than the other is the effect of the illusion). Going forward, I refer to the relevant cognitive constituents of an attitude as thoughts, leaving open the possibility that the constituent thoughts amount to beliefs.

Together with Characteristic Thought, Cognitive Conditions explains why the quality of will condition on moral blameworthiness should be analyzed along the lines of Countervailing Reasons. Characteristic Thought implies that blame involves a thought that is true only if the

Graham, “A Sketch of a Theory of Moral Blameworthiness”, *Philosophy and Phenomenological Research*, 2012; Gideon Rosen, “The Alethic Conception of Moral Responsibility”, *The Nature of Moral Responsibility*, 2015

agent expresses a morally objectionable quality of will in the sense given by Countervailing Reasons, and Cognitive Conditions states that an attitude is appropriate only if the thoughts it involves are true; it follows that some instance of behavior is blameworthy only if that behavior expresses a morally objectionable quality of will in the sense given by Countervailing Reasons.

The apparent complexity of Characteristic Thought may strike some as a problem for the view, since it must be plausible that the thoughts in question are had by any agent who experiences guilt, resentment, or indignation.⁴⁸ It is important to stress, however, that Characteristic Thought does not imply the absurd proposition that every instance of blame literally involves the thought “A’s beliefs and desires rationalize Xing, A possesses moral reasons M to refrain from Xing, and A’s motives for Xing do not counterbalance M”. Rather, it claims that it is a common feature of blame attitudes that these attitudes involve a thought that is true only if an agent meets the conditions given in Countervailing Reasons. When one considers the cases we have so far discussed, this claim should not seem implausible. For instance, it is not implausible that feeling indignation about Wanda’s lie in Weakness is partially constituted by a thought that Wanda’s motivations, her desire to claim victory, etc., do not justify the decision to lie. Nor is it implausible that when one feels indignation at Irene’s malicious testimony in Interview, this indignation involves thinking that Irene ought to base her evaluation of Janet on Janet’s merits as a candidate, something that she fails to do when she is motivated to testify against Janet out of malice. The content of such thoughts is best analyzed, I claim, along the lines of the conditions in Countervailing Reasons.

Of course, one does not consciously rehearse even these thoughts when one blames someone, but Characteristic Thought does not suggest that one does. Characteristic Thought is fully

⁴⁸ Rosen [2015] makes a similar point, arguing that the fact that children can blame constitutes a “naivety constraint” on theories of blame.

compatible with there being instances of blame where the only conscious elements are the affective elements of the emotion, the distinctive “pangs” of guilt, a sense of righteous anger if one is feeling resentment or indignation, etc. Even in cases where elements such as these are the only conscious elements of a blame emotion, there is good reason to think that the conscious elements reflect unconscious thoughts about the offending agent’s quality of will. Consider the following data.

- (A) We refrain from blaming when it is clear that an agent has not expressed a morally objectionable quality of will even when she causes us to suffer some harm or setback that would otherwise be cause for resentment. In cases where guilt, indignation, or resentment persists despite it being obvious that no objectionable quality of will was expressed, it is experientially clear that the emotion is recalcitrant, that is, that the attitude is not appropriately responsive to relevant information. The best explanation of why it is in this way transparent that blameworthiness requires the expression of an objectionable quality of will is that guilt, resentment, and indignation involve a thought that the act or omission expressed a morally objectionable quality of will, a thought that, in the case of recalcitrant blame, conflicts with one’s better judgment.
- (B) When we feel resentment or indignation, we are given to colorful expressions like “You did it to hurt me” or “You wanted me to fail”. These expressions of blame are so natural and immediate that they are best understood as literally giving voice to thoughts already implicit in the emotion. Thus, since the verbal expressions are manifestly about the morally objectionable quality of will expressed in an act or omission, they provide good evidence that blame involves a thought with similar content.

- (C) The fact that some behavior expresses a morally objectionable quality of will seems not only a constraint on the appropriateness of moral blame, but a reason for feeling resentful, indignant, or guilty.⁴⁹ A sensitivity to the fact that one has expressed an objectionable quality of will towards others seems part of what makes the emotion guilt different from mere regret, and part of what makes resentment and indignation have the character of “righteous” anger. The best explanation of why these reactive attitudes are not only caused by behavior that expresses a morally objectionable quality of will, but are in this way responses to the quality of will expressed, is that these attitudes involve the thought that the behavior in question expressed a morally objectionable quality of will.

I think that these considerations provide good reason to think that whatever form blame takes, whether it is guilt, resentment, indignation, or some other related emotion, it involves a thought that bears on the moral quality of the offending agent's will. Characteristic Thought is a hypothesis about what thoughts these attitudes have in common. Together with the fact that Characteristic Thought explains why Countervailing Reasons is a condition on the appropriateness of moral blame, I believe that this is good reason to think that Characteristic Thought is true.

2.3 Objections

In this section, I consider some objections to the claims made in sections 2.1 and 2.2 that clarify the view and some of its further consequences.

⁴⁹ On the other hand, while it is somewhat plausible that, e.g., blame is appropriate only if the target agent had the capacity to act otherwise, this seems more like a constraint on the appropriateness of blame rather than a reason for blaming. Whether there is any such capacity condition is discussed in the next chapter.

Objection: The claim that Characteristic Thought explains why the appropriateness of moral blame depends on the expression of a morally objectionable quality of will in the sense given by Countervailing Reasons assumes that the cognitive elements of an attitude have an independent meaning that contributes to whether the attitude is appropriate. In fact, the reverse is true: the meaning of a thought involved in an attitude is determined in concert with the other elements of the attitude.

This objection suggests that the views defended in section 2.2 assume a *cognitivist* theory of the blame attitudes and are incompatible with a *sentimentalist* theory of those attitudes. Cognitivists and sentimentalists disagree about whether the evaluative thoughts involved in blame and other attitudes contribute an independent meaning to the appropriateness of the attitude in question or whether the meaning of the thoughts is partly determined by the role the thought has as a constituent of the relevant attitude. For instance, a cognitivist about an emotion like fear would maintain that fear is an appropriate response to danger because it involves the thought *I'm in danger*. A sentimentalist would claim that the thought *I'm in danger* involved in fear gets its meaning in part by being part of an emotion that is an appropriate response to danger.⁵⁰

Whichever view is correct, however, fear is only appropriate if it is true that one is in danger. Neither Cognitive Conditions nor Characteristic Thought claims that the thought about quality of will involved in blame gets its meaning independently of being part of a blame attitude. The views defended above are neutral on order of explanation between the meaning of a constituent

⁵⁰ See D'Arms and Jacobson [2000] for a defense of a sentimentalist program. (Justin D'Arms and Daniel Jacobson, "Sentiment and Value", *Ethics*, 2000)

thought and the appropriateness of an emotion, and are therefore compatible with some sentimental theories.⁵¹

Objection: An agent's behavior can express a morally objectionable quality of will even if she does not possess any moral reasons not to engage in said behavior.

A case from Chapter One might be thought to support this objection.

Concrete: Carla is laying a foundation for a house and in the course of some calculations comes to the incredible conclusion that, given the slope of the land, the soil conditions, and various other features of the project, without reinforcement, an ordinary concrete foundation will become unstable and the house will collapse. Carla is initially shocked by her findings, then realizes that this is the perfect opportunity to kill her despised sister, Dolores, who will occupy the house. Intent on killing Dolores, Carla does not reinforce the foundation.

In fact, Carla is right to find it incredible that the house will collapse unless the foundation is given additional reinforcements: Carla's conclusion was the result of a gross error in her calculations, and no evidence she has about the project conditions makes the conclusion even remotely plausible.

However, as the result of a dramatic shift in weather patterns, a gradual rise in the water table, changes in local soil composition, and erosion, the foundation of Dolores's house

⁵¹ I in fact favor the view that the meaning of the thought about quality of will involved in blame is partly determined by the non-cognitive elements of blame adduced in Chapter Three.

becomes unstable and the house collapses, killing Dolores. Had Carla reinforced the foundation, the house would not have collapsed.

Carla expresses a morally objectionable quality of will when she decides not reinforce the foundation. According to Countervailing Reasons, then, Carla must have some moral reason to decide to reinforce the foundation. However, it is natural to think that the only reason to reinforce the foundation — that it would prevent Delores being crushed — is not a reason that Carla possesses, since Carla has no good evidence that Delores will be crushed.

On my view, Carla does possess a moral reason to reinforce the foundation: the fact that it seems to Carla that failing to reinforce the foundation will result in Delores's death. Of course, this reason is defeated by the overwhelming evidence there is that reinforcement is not necessary to prevent Delores' death, but it is a reason nonetheless. We should not infer from the fact that a reason has little or no weight that it is not a reason.⁵²

Objection: If evidence of a reason is a reason, as the treatment of Concrete here suggests it is, then it is possible for an agent to express no deficient quality of will even if she possesses moral reasons that count decisively against an act, since an agent can possess evidence and yet not grasp its significance.

The following case might be thought to provide further support to the objection.

Nuclear: Several workers are trapped in a nuclear power plant. Nancy knows that if a certain conjecture in nuclear physics is true, she must turn off the reactor by midnight or

⁵² Schroeder [2007] argues persuasively that negative existential claims about reasons are often unreliable. (Mark Schroeder, *Slaves of the Passions*, Oxford University Press, 2007)

it will explode, killing the workers. On the other hand, if the conjecture is false, turning the reactor off will cause it to explode. Nancy knows the truth or falsity of the conjecture can be derived from a set of data that she has, but despite her best effort to derive the answer, she does not succeed in doing so before midnight arrives. The reactor explodes.

Nancy is aware of facts that constitute decisive moral reason to shut down the reactor, namely, the facts that entail that if she doesn't do so, the workers will be killed. The objection maintains that since her failure does not express a morally objectionable quality of will, Countervailing Reasons must be false.

On my view, while the evidence of which Nancy is aware does constitute decisive moral reason to shut down the reactor, Nancy does not possess the moral reasons constituted by that evidence. In Concrete, Carla has evidence that her actions might have a morally bad outcome (i.e., Delores's death) and, further, is sensitive to that evidence: the apparent result of her calculations moves her to hold a belief (unreasonably strong though it is) that failing to reinforce the foundation will result in Delores's death. Nancy, on the other hand, is not sensitive to the evidence that failing to shut down the reactor will result in the death of the workers: although she knows that shutting down the reactor will either cause a deadly explosion or will prevent a deadly explosion, she has no idea which it is. To possess a moral reason — or a practical reason more generally — an agent must be sensitive to the properties of that reason that make it a reason. Since Nancy is not sensitive to the significance of her evidence, she does not, as the objection assumes, possess decisive moral reason to shut down the reactor. Countervailing Reasons thus correctly implies that Nancy does not express a morally objectionable quality of will.

Objection: An agent who acts wrongly in a state of self-deception may not possess the moral reasons that make it wrong to act as she does but nonetheless express a morally objectionable quality of will. Countervailing Reasons falsely implies that such agents are blameless.

The following case might be thought to provide further support to the objection.

Snake Oil: Sasha travels around her country making and selling snake oil. She sells the snake oil as a cure for a wide range of maladies, charging prices that deplete the savings of her desperate customers. Sasha is careful to avoid returning to towns she has visited in the past; when she does find herself in a community she has visited before, past customers are angry and call her “quack”. Influenced by her desire to believe in the powers of her snake oil and to continue her business, Sasha downplays these claims in her mind and continues to believe that the snake oil really does work in many cases. If this belief were true, Sasha’s high prices for the snake oil would be justified. In fact, however, the snake oil is medically useless.

Clearly there is a moral reason for Sasha not to sell the snake oil — the fact that it does not have the benefits that Sasha advertises or that the buyers believe that it has — but given the conviction Sasha has in her product’s powers, she does not possess this reason. According to Countervailing Reasons, then, her actions do not express a morally objectionable quality of will. This result, however, is unacceptable: Sasha’s behavior clearly expresses too little concern for the rights and interests of her customers.

On my view, Sasha is blameworthy because her behavior involves serious mismanagement of her beliefs. The fact that Sasha avoids returning to towns she has visited in the past reveals, if

not a suspicion that snake oil lacks the medicinal powers she advertises, at least an intention to avoid evidence that might test her conviction in her snake oil's powers.⁵³ Further, when faced with insults and accusations that the snake oil is impotent, Sasha could reflect on this evidence, but is motivated to avoid doing so. Intentionally avoiding evidence that bears on the efficacy of her snake oil is morally objectionable, especially given that Sasha knows that many of her customers impoverish themselves to buy it. I thus claim (i) that Sasha's beliefs and desires rationalize her behavior, (ii) that her epistemic state includes moral reasons to manage her beliefs about the snake oil differently, and (iii) that her motivations — to avoid confronting unpleasant facts, etc. — do not defeat or outweigh those reasons. Sasha's mismanagement of her beliefs satisfies the conditions for expressing a morally objectionable quality of will given by Countervailing Reasons.

Objection: It is plausible that in Kitten, Kate is blameless for failing to shout out, even though she is capable of doing so. However, it is possible for there to be cases that are structurally parallel to Kitten — cases in which an agent is aware of reasons to act a certain way, has beliefs and desires that rationalize so acting, and is capable of acting on those beliefs and desires — but that involve not a morally wrong omission like Kate's, but a morally wrong action. In these cases, it is implausible that the agent is blameless.

This objection holds that there are counterexamples to the claim that, in order to be blameworthy for an act, an agent's beliefs and desires must rationalize her behavior. These are cases in which (i) an agent is aware of moral reasons that require her to refrain from a certain course of action, (ii) the agent is capable of refraining from that course of action, (iii) the agent's beliefs and desires rationalize refraining from that course of action, but (iv) the agent performs

⁵³ Sasha's self-deception is in this way different from the naïve belief mismanagement that occurs in cases like Botulism, discussed in Chapter One.

the action nonetheless. But are such cases really possible? And if they are, are the agents involved blameworthy?

One case that may support a positive answer to the first question is Stanley Milgram's famous obedience experiment. As described in Milgram [1983], the subjects in this experiment volunteered to participate in what they believed to be a study on learning and memory.⁵⁴ They were assigned to fill a role as a "teacher", which required the subject to ask a series of word association questions through an intercom to a "learner" in another room and to deliver progressively stronger shocks in 15-volt increments when the learner answered incorrectly. The subjects of the experiment believed that the learner was a volunteer like themselves, strapped to a machine that would deliver the electric shocks; in fact, the learner was a confederate of the experiment who received no shocks and responded by playing a recording. The recording first played grunts of pain, then loud complaints, then demands to be released from the experiment on the basis of a heart condition, and finally, after one final scream, total silence. Despite what seems to be overwhelming moral reason to cease the experiment, 80% of subjects delivered shocks past the point where the learner explicitly withdrew consent; 63% continued to deliver shocks until the experimenter instructed them to stop, well after the point where the learner fell silent. They continued only because the experimenter calmly insisted that the experiment "must go on", that the shocks were painful, but not dangerous.

I will assume (i) that Milgram's subjects were aware of moral reasons that required them to refrain from delivering the shocks, and (ii) that they were capable of refraining, capable of refusing the experimenter's demands to go on. We know (iv) that many subjects failed to resist these demands. The remaining question, then, is whether the decision to continue delivering the shocks was rationalized by the subjects' preferences.

⁵⁴ Stanley Milgram, *Obedience to Authority*, Harper Perennial, 1983.

According to Arpaly and Schroeder [2013], the subjects of the Milgram experiment acted contrary to their preferences.⁵⁵ Their primary piece of evidence is a later variation of Milgram's original experiment in which subjects first observed someone else refuse the experimenter's instructions to deliver a shock, and were then asked to deliver the shocks instead. In this version of the experiment, many more subjects refused than did in the original version. Arpaly and Schroeder argue as follows.

That their refusal needed such a weak trigger — essentially, just a reminder that it could be done — strongly suggests that many participants were acting contrary to what they on balance desired in seemingly delivering the harmful shocks. (Arpaly and Schroeder [2013], 168)

Arpaly and Schroeder find it implausible that the subjects have desires that constitute a kind of conditional preference, *If no one else has refused the experimenter's instructions, then shock; if someone else has refused the experimenter's instructions, then don't shock*. Rather, they maintain that Milgram's subjects had beliefs and desires that rationalized ceasing the experiment all along, but that the conditions of the experiment caused them to fail to act in accordance with this preference.

This is not the only plausible assessment of the subjects' behavior. For instance, it seems equally possible that, e.g., prior to seeing someone else refuse to obey the experimenter, the subjects were uncertain about whether they ought to obey or disobey and that delivering the shocks was rationalized by a strong desire to obey a legitimate authority; seeing someone else disobey the experimenter could then have facilitated disobedience by causing the subject to

⁵⁵ Nomy Arpaly and Timothy Schroeder, *In Praise of Desire*, Oxford University Press, 2013.

realize that the experimenter's demands were unreasonable. However, the objection only requires that it is possible that the subjects' preferences did not rationalize delivering the shocks. Let us assume that it is possible, and stipulate (iii) that the subjects' preferences did not rationalize delivering the shocks, and in fact rationalized refusing to do so, but that they delivered the shocks nonetheless. If Countervailing Reasons is true, then the subjects are blameless. The objection holds that this is false, that the subjects are blameworthy for taking actions that they believed would deliver powerful shocks to an unconsenting victim.

Cases such as this, however, are not ones where an intuition that an agent is or is not blameworthy should be a guide to theory. Rather, given the numerous stipulations that are required to isolate the interpretation of the Milgram subjects' behavior that is relevant to the objection, judgments about whether these agents are blameworthy should be guided by the theory of blameworthiness that we have developed by considering numerous other, more intuitive cases. And according to the view developed above, if the Milgram subjects' desires did not rationalize their behavior, then, morally speaking, that behavior is like someone's freezing with panic or acting on an addictive impulse despite the fact she would prefer to act otherwise: such behavior does not express an agent's will, and so is not appropriate cause for moral blame. Whether any of Milgram's actual subjects have such an excuse is, of course, a further question.

3. Blameworthiness, Sanctions, and Desert

This chapter argues that any moral norm that bears on whether an agent merits sanctions for her behavior also bears on whether she is blameworthy for that behavior. The chapter answers recent criticism of this position by identifying a feature of moral blame that explains why blame should be governed by the same moral norms that determine whether an agent merits sanctions, namely, that in blaming someone for engaging in some behavior, one takes there to be no backward-looking moral principles that prohibit sanctioning her for that behavior. This discussion motivates a novel account of blame that contrasts with pure cognitivist theories of blame and theories on which blame is essentially retributive.

No moral theory is complete without a discussion of *moral sanctions*, forms of harsh treatment imposed on an agent as a means of condemning some objectionable behavior in which she has engaged. Although there is a general moral obligation to avoid treating agents in ways that they will find disagreeable or unpleasant, the fact that an agent's behavior constitutes a certain kind of moral offense can make for an exception to this prohibition such that it is *prima facie* permissible to sanction her. Stipulate that in such cases the agent *merits* sanctions for her behavior.⁵⁶

It is widely agreed that in order for an agent to merit sanctions for some behavior, she must be *blameworthy* for engaging in that behavior, that is, it must be appropriate for the offending agent

⁵⁶ That a sanction is merited contributes to its being morally justified, all things considered, but it is neither a necessary nor a sufficient condition for justification. There may be good reasons not to sanction an agent, even if her behavior does in fact merit sanctions. Similarly, although it is generally forbidden to sanction the innocent, it is possible for there to be overriding considerations such that imposing sanctions is morally justified, all things considered.

to feel guilty about this behavior and appropriate for others to feel indignant or resentful. Call this the thesis that Meriting Sanctions is Conditional on Blameworthiness.

Meriting Sanctions is Conditional on Blameworthiness: An agent A merits sanctions for engaging in behavior X only if A is blameworthy for X.

In this chapter, I argue for the converse of this thesis, the position that it is only appropriate to blame an agent for her behavior if that behavior merits sanctions. Call this the view that Blameworthiness is Conditional on Meriting Sanctions (BCMS).

Blameworthiness is Conditional on Meriting Sanctions: An agent A is blameworthy for behavior X only if A merits sanctions for engaging in X.

Several theorists have made claims that are suggestive of BCMS. For instance, Wallace [1994] argues that in virtue of the fact that blaming an agent disposes one to sanction her, an agent is blameworthy only for behavior over which she has a certain kind of control.⁵⁷

[I]t would be unfair to hold someone to moral obligations one accepts [by blaming her] in the absence of the powers of reflective self-control [because] it is unreasonable to demand that people do something — in a way that potentially exposes them to the harms of moral sanction — if they lack the general power to grasp and comply with the reasons that support the demand. (Wallace [1994], 161)

Rosen [2003] supports Wallace's claim, arguing that blame is itself a kind of sanction.⁵⁸

⁵⁷ R. Jay Wallace, *Responsibility and the Moral Sentiments*, Harvard University Press, 1994.

⁵⁸ Gideon Rosen, "Culpability and Ignorance", *Proceedings of the Aristotelian Society*, 2003.

As is commonly observed, moral blame is a sort of sanction. Even when it is not expressed, it is a form of adverse treatment: a form of psychic punishment. And like any adverse treatment, blame is governed by moral norms. (Rosen [2003], 73)

If the supporters of BCMS are correct, then blame is governed by all of the same moral norms that bear on whether an agent merits sanctions for her behavior.

Recently, such views have received significant criticism. For instance, Hieronymi [2004] argues that the appropriateness of moral blame does not depend on the satisfaction of the same norms of fairness that govern sanctions because, whereas sanctions are unpleasant and burdensome, unexpressed blame is not.⁵⁹ Similarly, Graham [2012] argues that since being disposed to treat an agent in a certain way does not involve actually treating her in that way, there is no obvious reason that attitudes of moral blame should be constrained by the norms that govern sanctions and other forms of harsh treatment.⁶⁰

In the first section of this chapter, I present a novel argument for the claim that Blameworthiness is Conditional on Meriting Sanctions. The key premise of my argument is that attitudes of moral blame do not merely dispose one to sanction an offending agent for her behavior, but further involve taking there to be no moral prohibitions of a certain kind on the imposition of sanctions. Agents are only blameworthy, I argue, if it is true that there are no such prohibitions.

In the second section of the chapter, I argue that taking there to be no relevant moral prohibitions on the imposition of sanctions is one of two characteristic features of blame. The

⁵⁹ Pamela Hieronymi, "The Force and Fairness of Blame", *Philosophical Perspectives*, 2004.

⁶⁰ Peter Graham, "A Sketch of a Theory of Moral Blameworthiness", *Philosophy and Phenomenological Research*, 2012.

second feature, on my view, is one that is amenable to philosophers on both sides of the debate over BCMS: a cognitive state that concerns the morally objectionable attitudes expressed by the offending agent's behavior. I defend this characterization of blame over alternatives, in particular, alternatives on which blaming an agent involves taking there to be a non-instrumental reason that counts in favor of sanctioning her. I argue that my preferred account has a number of theoretical advantages over such alternatives. For instance, in the debate over whether it is ever appropriate to blame an agent for behavior that is causally determined by facts outside of her control, I argue that my view has the advantage of putting the burden of proof on the *incompatibilist*, who gives a negative answer to the question, rather than on her *compatibilist* opponent. This, I claim, is where the burden of proof belongs.

In the third section, I consider some objections to the view, including an objection that presses the case for including a desire to sanction in the characterization of blame, an objection that the characterization of blame that I prefer does not account for the negative character of blame, and an objection that holds that the view is vulnerable to certain incompatibilist arguments that do not rely on the assumption that there are non-instrumental reasons that count in favor of imposing sanctions. These are by no means the only possible objections to the view, but each identifies important consequences of the view or other issues worthy of further consideration.

3.1 Sanctions and the Nature of Blame

For the claim that Blameworthiness is Conditional on Meriting Sanctions (BCMS) to be plausible, we must distinguish the question of whether sanctions are merited from the question of whether sanctions are morally justified, all things considered. This is essential because the fact that an agent is blameworthy for some act does not entail that it is permissible, all things considered, to sanction her. Consider a case.

Pumpkin: Patty makes the free and informed decision to steal Quinn's best pumpkin to enter in the local gardening competition. Nothing supports this decision other than Patty's desire to win and the fact that Quinn has the best pumpkin around. Quinn learns of Patty's actions on the day of the competition and has the opportunity to disqualify her. However, Quinn knows that Patty has a fragile disposition and will suffer a heart attack if she is disqualified or feels she is being sanctioned in any way. Quinn therefore refrains from disqualifying Patty, though she resents her greatly.

Patty is blameworthy for stealing Quinn's pumpkin. However, it would be wrong for Quinn to sanction Patty since doing so would cause her to suffer a heart attack.

In the case of Pumpkin, Quinn's obligation not to sanction Patty is grounded in *forward-looking* considerations, that is, the downstream effects that would follow if Quinn were to sanction Patty.⁶¹ However, whether an agent is blameworthy for engaging in some behavior depends only on *backward-looking* considerations, considerations that concern only the nature of the behavior in question and the way the agent came to engage in that behavior. Similarly, the prohibitions on moral sanctions that determine whether an agent merits blame are backward-looking. Consider, for instance, this principle.

Control (Sanctions): An agent does not merit sanctions for behavior that she lacks the rational capacity to control.⁶²

⁶¹ Smart [1961] famously defends a forward-looking conception of blameworthiness as a replacement for the norms that govern the ordinary backward-looking practice. (J.J.C. Smart, "Free Will, Praise and Blame", *Mind*, 1961)

⁶² A better version of this principle — and various similar principles I discuss in the chapter — will hold that agents do not merit sanctions for behavior that they *blamelessly* lack the rational capacity to control, or that they do not merit sanctions for such behavior *non-derivatively*, since one still might

Control (Sanctions) is a backward-looking principle in the sense that, according to this principle, whether an agent merits sanctions depends on the state the agent was in leading up to the behavior in question, namely, whether she had the rational capacity to control said behavior.

I will assume that the question of whether an agent merits sanctions for some behavior reduces to the question of whether there is a backward-looking moral principle that generally prohibits anyone from sanctioning the agent for said behavior. (A positive answer to the question of whether the agent merits sanctions entails a negative answer to the question of whether there is any relevant backward-looking prohibition on sanctions and *vice versa*.) Precisely what backward-looking moral principles there are is a substantive issue in ethics, but we do not need to resolve that issue here; what is important for our purposes is that if an agent does not merit sanctions, there is some backward-looking moral principle that explains why. This is the first premise of my argument for BCMS.

(P1) If an agent A does not merit sanctions for behavior X, then there is some backward-looking moral principle that prohibits sanctioning A for X.

This principle is uncontroversial. Having stated it, we are now in a position to turn our attention to moral blame.

On my view, blame attitudes play an intimate role in motivating people to impose sanctions on agents who engage in a certain kind of morally objectionable behavior. Consider a case.

merit sanctions if one is blameworthy for one's lack of control (e.g., if one's lack of control is the result of willful intoxication). I suppress this detail for ease of exposition.

Agreement: Abigail has agreed to sell her house to Beatrice. Beatrice doesn't have the money for a down payment, but she has a good job and will have the money soon. The two agree that Abigail will take the house off the market and that Beatrice will make the down payment in two weeks.

The agreed date passes and Abigail sends Beatrice a note. Days go by without a reply, so Abigail calls a real estate broker to ask for advice. The broker informs Abigail that records show that Beatrice viewed several more houses during the two weeks that Abigail's house was off the market and settled on one the previous week.

Upon learning that Beatrice broke her commitment to buy her house in this way, Abigail is overwhelmed with resentment and is moved to call Beatrice. Abigail normally has strong moral inhibitions to treating others harshly, but given Beatrice's behavior, Abigail does not now feel so constrained. When Beatrice answers, Abigail's resentment moves her to criticize Beatrice's actions as a "betrayal", as "cowardly" and "disrespectful". She then hangs up.

Abigail's resentment is a form of blame. And like many instances of blame, Abigail's resentment disposes her to sanction Beatrice, that is, it makes it more likely that she will take the steps to deliver the harsh criticism indicated above. How does Abigail's resentment dispose her to sanction Beatrice? It is possible for blame to play its role behind the scenes, setting the stage for an agent to decide to impose sanctions, but that is not how things progress in Abigail's case: Abigail's resentment involves her taking Beatrice's behavior to be a reason that supports her decision to impose sanctions on Beatrice.

The state of *taking something to be a reason* is a practical state, distinct from simply judging that something is a reason. One can judge that something is a reason without taking it to be a reason, and *vice versa*. For instance, jealousy might cause one to take a friend's success to be a reason to avoid her, even though one is confident in the contrary judgment that her success in no way constitutes a reason to behave this way. On the other hand, one might judge that the suffering of animals is a reason to refrain from eating meat, but fail to take this consideration to be a reason that bears on one's own decision of what to order at a restaurant. Abigail does not explicitly judge that Beatrice's behavior is a reason to sanction her, but she nonetheless takes it to be a reason to do so.

Part of the reason that Abigail's resentment is manifested in her sanctioning Beatrice by calling her and delivering harsh criticism is that Abigail has some desire that is satisfied by her doing so, a desire to condemn such behavior, perhaps, or just a desire for vengeance. However, Abigail does not take Beatrice's behavior to be a reason for sanctioning her just because it is a means to satisfying this desire; the kind of resentment or indignation that constitutes moral blame is not just a form of anger, it is a form of *righteous anger*, anger that presents itself as warranted or called for. On my view, what makes Abigail's resentment a form of righteous anger is that part of what is involved in Abigail's resentment, and in blame in general, is taking there to be no backward-looking moral prohibitions on sanctioning the offending agent for the behavior in question. This is the second premise of my argument.

- (P2) Blaming an agent A for some behavior X involves taking there to be no backward-looking moral principle that prohibits sanctioning A for X.

The fact that blaming someone involves taking her behavior to support the imposition of sanctions in this way explains not only the righteous quality of resentment and indignation, but

also some of the unpleasant phenomena associated with guilt, the paradigm form of self-blame. Genuine guilt, on my view, involves feeling that one lacks a certain moral protection against sanctions. Most dramatically, guilt can be manifested in the form of penance or self-imposed sanctions. Of course, it is also possible to both feel guilty and to have a sincere desire to avoid sanctions of any sort; even in this latter sort of case, however, a guilty agent takes her behavior to provide others a kind of license to sanction her. This is why, while a guilty agent might hope for forgiveness, she has a sense of being at others' mercy.⁶³

Why think that blame involves taking there to be no moral prohibitions on sanctioning an agent for her behavior? Isn't it more plausible that the agent's behavior defeats just one such prohibition, the general prohibition on harsh treatment that is always defeated when an agent engages in some blameworthy behavior? There are, I think, several reasons to prefer my view. First, the sense of being at others' mercy that is involved in guilt only makes sense if one merits sanctions; in cases where one does not merit sanctions, others do not have the kind of license to impose sanctions that warrants feeling that one is at their mercy. The same is true when one blames others: the moral authority that one feels one has when one experiences resentment or indignation suggests that one has a kind of license to impose reasonable sanctions.

In a somewhat different vein, there is a serious tension between, on the one hand, a confident belief that an agent does not merit sanctions and, on the other, wholeheartedly blaming her. On the present view, this tension is explained by the fact that blaming involves taking there to be no backward-looking prohibitions on sanctions, which is unreasonable when one sincerely believes

⁶³ My view of blame sits well with theories like that articulated by Calhoun [1992], on which forgiveness is "elected" rather than owed to an offender (Cheshire Calhoun, "Changing One's Heart", *Ethics*, 1992). However, it is still possible for an agent to deserve forgiveness, e.g., because of steps the offending agent takes subsequently to make up for her prior objectionable behavior.

the contrary.⁶⁴ For instance, if it is clear that an agent does not merit sanctions because, e.g., she acted under extreme duress, or involuntary intoxication, then it seems unreasonable to persist in blaming her. This is not the case when blame is prohibited for purely forward-looking reasons; in cases like Pumpkin, Patty is clearly blameworthy even though her fragile disposition greatly outweighs whatever reasons Quinn might have to disqualify her from the competition.

The final step in my argument requires a thesis that connects the state of taking there to be a certain reason or prohibition to the appropriateness of an attitude like moral blame. I endorse the following principle.

Conative Conditions: If an attitude A involves taking a fact F to bear on whether one should Φ in virtue of norm N, then A is appropriate only if F does bear on whether one should Φ in virtue of norm N.

Conative Conditions is a complex principle. I will first support it by illustrating that blame is plausibly governed by a simpler but analogous principle, and then argue that Conative Conditions has independent plausibility.

The debate over whether any principle like BCMS is true is an intramural debate between theorists who, influenced by Strawson [1962], understand blame to be a reaction to behavior that seems to express a *morally objectionable quality of will*, a lack of respect for the rights and interests of others.⁶⁵ What makes blame a “reaction” to such behavior? The standard view now is that it is a certain *cognitive state* involved in blame, a thought, belief, or judgment to the effect that the offending agent’s behavior expresses a morally objectionable quality of will. Hieronymi

⁶⁴ Of course, it might not be one’s inclination to blame that is mistaken, but rather the belief that the agent is blameless.

⁶⁵ Peter F. Strawson, “Freedom and Resentment”, *Proceedings of the British Academy*, 1962.

[2004] and Graham [2012] explicitly endorse such a view, maintaining that blame is appropriate only if this thought, belief, or judgment is accurate.⁶⁶

Cognitive Conditions: If an attitude A involves a cognitive element with content P, then A is appropriate only if P is true.

The case for accepting Cognitive Conditions is very strong. Together with the hypothesis that blame involves a thought that implies that the offending agent has expressed a morally objectionable quality of will, it explains why a wide variety of agents are not blameworthy. For instance, while Abigail is correct that Beatrice expresses a morally objectionable quality of will in Agreement when Beatrice breaks her commitment to buy Abigail's house, Abigail might have been mistaken: It might have been the case that Abigail's real estate broker confused Beatrice with another buyer and that the real reason Beatrice did not make her payment was not that she was looking at other houses, but because she suffered a stroke. In this case, Beatrice would not be blameworthy for failing to pay Abigail because this failure would have expressed no morally objectionable quality of will. Anyone blaming Beatrice, including Abigail, would have a false thought about Beatrice's quality of will.

The analogy between Conative Conditions and Cognitive Conditions is this: If our claim about blame were not that it involves taking there to be no backward-looking moral prohibition on imposing sanctions, but rather thinking that there is no such prohibition, then, by Cognitive Conditions, it would follow that blame is appropriate only if it is true that there is no backward-looking moral prohibition on the imposition of sanctions. Conative Conditions implies that the

⁶⁶ Graham [2012] ultimately argues that the proper locus of blame is not an agent's behavior, but the mental states that violate the relevant moral norms of respect. However, like the argument offered by Hieronymi [2004], his argument emphasizes the role blame's cognitive constituents have in determining when blame is appropriate.

state of taking there to be no relevant prohibitions has the same effect on appropriateness as would a thought to this effect: only if an agent's behavior (F) really does bear on whether one should sanction her (Φ) by being the kind of behavior for which backward-looking moral principles do not prohibit sanctions (N) is it appropriate to take her behavior to have this property.

In addition to this analogy with Cognitive Conditions, Conative Conditions can be directly supported with cases of hostile emotions that do not involve the same thoughts about quality of will that are involved in blame. Consider a case.

Classroom: Charlene is enjoying having a large and convenient desk to herself in one of her classes when a new student, Deborah, joins the class and is instructed to next to Charlene. Charlene immediately forms a hostile attitude towards Deborah: Not only is Charlene averse to offering Deborah even the common courtesies that would be normal in their circumstances, she actively hopes that Deborah will do poorly in the class. Charlene knows that Deborah does not bear her any ill will; her hostility is a reaction to the sheer inconvenience of sitting next to Deborah.

Charlene's hostility towards Deborah is clearly inappropriate and undeserved. If the hostility Charlene feels towards Deborah were a form of blame, one could explain its inappropriateness in terms of Cognitive Conditions and the falsity of the thought that Deborah's behavior expresses a morally objectionable quality of will. But Charlene's hostility is not a form of blame; it is a reaction to the inconvenience of sitting next to Deborah, not to Deborah's quality of will. And as a matter of fact Deborah's sitting at Charlene's desk is inconvenient for Charlene. It is thus not clear that we can attribute the inappropriateness of Charlene's attitude to any mistaken cognitive constituent.

What makes Charlene's attitude towards Deborah inappropriate, on my view, is that it involves taking the fact of Deborah's sitting at her desk to be a reason to treat Deborah with hostility when, in fact, Deborah's sitting at her desk gives Charlene no such reason. Of course, it is conceivable that Charlene has an *instrumental* reason to be hostile to Deborah — as would be the case if, for instance, being hostile would enable Charlene to recover her valuable desk space — but the natural interpretation of Charlene's hostility is not so calculating. Rather, Charlene takes Deborah's sitting at her desk to be a *non-instrumental* reason for being hostile to Deborah, a reason that supports being hostile irrespective of any downstream effect that the hostility will have. According to Conative Conditions, then, Charlene's hostility to Deborah is appropriate only if Deborah's sitting at Charlene's desk (F) supports being hostile towards Deborah (Φ) in virtue of being a non-instrumental reason to do so (N). Since Deborah's sitting at Charlene's desk is not a non-instrumental reason for Charlene to be hostile towards Deborah, it follows that Charlene's hostility is inappropriate. That Conative Conditions can explain the inappropriateness of such cases is good evidence that it is true.

Conative Conditions supports the final premise (P3) in my argument for BCJS.

- (P3) If blaming an agent A for behavior X involves taking there to be no backward-looking moral principle that prohibits sanctioning A for X, then A is blameworthy for X only if there is no backward-looking moral principle that prohibits sanctioning A for X.

Again, P3 is an instance of Conative Conditions where the offending agent's behavior (F) is taken to support the imposition of sanctions (Φ) in virtue of being the kind of behavior for which backward-looking moral principles do not prohibit sanctions (N).

We can now state the argument for BCMS.

The Argument that Blameworthiness is Conditional on Meriting Sanctions:

- (P1) If an agent A does not merit sanctions for behavior X, then there is some backward-looking moral principle that prohibits sanctioning A for X.
- (P2) Blaming an agent A for some behavior X involves taking there to be no backward-looking moral principle that prohibits sanctioning A for X.
- (P3) If blaming an agent A for behavior X involves taking there to be no backward-looking moral principle that prohibits sanctioning A for X, then A is blameworthy for X only if there is no backward-looking moral principle that prohibits sanctioning A for X.
- (C1) An agent A is blameworthy for behavior X only if there is no backward-looking moral principle that prohibits sanctioning A for X. (From P2, P3)
- (C2) (BCMS) An agent A is blameworthy for behavior X only if A merits sanctions for engaging in X. (From P1, C1)

To reiterate: P1 is an account of when sanctions are merited. P2 is an account of a fundamental feature of blame that explains the righteous character of resentment and indignation, the vulnerability involved in guilt, and various other phenomena. P3 is an instance of Conative Conditions. If this argument is successful — as I believe it is — then BCMS is true.

3.2 Blameworthiness and Desert

On my view, the discussion in the previous section completely characterizes what unifies resentment, indignation, and guilt and makes these attitudes forms of blame. These attitudes

are reactions to behavior that expresses a morally objectionable quality of will and which are sensitive to the significance such behavior has for the justification of sanctions. More precisely, I endorse the following characterization of moral blame.

Blame Attitudes: For an agent A to blame an agent B for behavior X is for A to have an attitude E that involves:

- (i) thinking that B's performance of X has features that, as a matter of fact, are sufficient for X to express a morally objectionable quality of will, and
- (ii) taking there to be no backward-looking moral principle that prohibits sanctioning A for X.

As noted above, particular blame emotions may have idiosyncratic features that are not common to other blame emotions. (Although blame does not necessarily involve a desire to sanction the offending agent, for instance, Abigail's resentment involves a desire to sanction Beatrice that ultimately moves her to call Beatrice and to deliver criticism in a way that would be impermissible if Beatrice did not merit sanctions.) The two features given in Blame Attitudes, on the other hand, are common to every instance of blame. On my view, it is these features that make a particular attitude a form of blame.

The philosophical literature has recently seen a great deal of disagreement over how exactly moral blame should be characterized. It has been argued that blame involves a kind of protest of a blameworthy act (Smith [2012]), a desire that a blameworthy act not been performed (Sher [2005]), a judgment that a blameworthy act counts in favor of decreasing or limiting one's relationship with the agent who performed it (Scanlon [2008]), and so on.⁶⁷ It is sometimes

⁶⁷ Angela Smith, "Moral Blame and Moral Protest", *Blame: It's Nature and Norms*, 2012.
George Sher, *In Praise of Blame*, Oxford University Press, 2005.

unclear whether these theories are, on the one hand, intended as factual claims about some set of actual mental states, e.g., guilt, resentment, and indignation, or on the other hand, whether these theories offer a kind of recommendation about how philosophers should use the term “blame”, given various issues that are of interest to philosophers in discussions of blame and blameworthiness.

I have given reasons for thinking that Blame Attitudes gives an account of the elements that are common to actual instances of guilt, resentment, and indignation. That blame involves taking an offending agent’s behavior to merit sanctions explains much about the way in which guilt, resentment, and indignation are experienced. Because Blame Attitudes does not include a desire to impose sanctions, it correctly predicts that there are instances of guilt that are not self-flagellating, that is, that while painful in the way that guilt is always painful, do not involve a desire for suffering. Similarly, although resentment and indignation often involve a desire to sanction or criticize, it allows for the possibility of “cool” instances of these emotions that do not involve such a desire. It is a consequence of Blame Attitudes that even if the cool analogue to resentment is not a typical form of resentment, it is still a form of blame.

One might worry that this goes too far, that we should draw the circle of what counts as blame smaller than does Blame Attitudes. One suggestion with some initial plausibility is that blame involves not only taking the offending agent’s behavior to merit sanctions, in the backward-looking sense, but to *count in favor* of sanctions, that is, to weigh positively in the calculation as to whether one ought to impose sanctions, all things considered. The most interesting version of this suggestion claims that blame necessarily involves treating an agent’s behavior as a non-instrumental reason that counts in favor of sanctions.

Deserved Sanctions: Blaming an agent A for behavior X necessarily involves taking A's engaging in X to be a non-instrumental reason that counts in favor of sanctioning A for X.

Blame Attitudes suggests that feeling blame disposes one to sanction an offending agent insofar as it silences any inhibitions one has to imposing sanctions that are sensitive to backward-looking moral prohibitions on imposing sanctions. In this way, blame will dispose one to sanction offending agents just in case one takes there to be good forward-looking reasons to do so. According to Deserved Sanctions, however, no forward-looking reason need be identified: the very fact that the agent engaged in the behavior in question is a reason to sanction her.⁶⁸

The difference between Blame Attitudes and Deserved Sanctions can be illustrated with the case of Agreement. On my view, Abigail's resentment of Beatrice involves judgments about the morally objectionable nature of Beatrice's behavior, Abigail's taking there to be no backward-looking moral prohibitions on sanctioning Beatrice for this behavior, plus a desire to criticize Beatrice for this morally objectionable behavior. For the sake of argument, let us assume that this desire of Abigail's is an intrinsic desire to express her negative judgments about Beatrice's behavior, and that the harshness of the words Abigail says to Beatrice is a means to emphasizing these judgments. On my view, part of the reason Abigail acts on this desire is that she takes herself to have the license to do so, since she takes there to be no backward-looking prohibitions on sanctioning Beatrice in this way.

⁶⁸ Pereboom [2014] argues that blame involves a belief that the offending agent *basically deserves*, deserves in a fundamental sense, to be the recipient of expressions of moral blame. On my view, Deserved Sanctions is a more plausible version of Pereboom's view, since an agent need not have a belief that must be understood in terms of an esoteric philosophical concept, but only take there to be non-instrumental reasons to impose sanctions. (Derek Pereboom, *Free Will, Agency, and Meaning in Life*, Oxford University Press, 2014)

Deserved Sanctions adds to this story. Abigail not only has an intrinsic desire to emphasize her negative judgments about Beatrice's behavior, she takes the fact that Beatrice acted as she did to itself count in favor of sanctioning her. In this version of the case, Abigail feels that Beatrice deserves to be criticized harshly in this way and is moved in part by this conviction. On my view, while these do not make an implausible addition to Abigail's case, it seems to me also quite possible that such elements are absent.

On the question of what characterization of blame is most fruitful for philosophy, I believe that there is reason to prefer Blame Attitudes as it stands. One reason concerns the relationship between blame and punishment. Feinberg [1965] argues that what makes a sanction a form of punishment rather than a simple penalty is that it expresses an attitude of moral condemnation.⁶⁹ On my view, the moral condemnation at issue is that the offending agent expressed a morally objectionable quality of will in engaging in the offending behavior. Since sanctions motivated by attitudes that conform to Blame Attitudes serve to express this thought even if they do not have the features attributed in Deserved Sanctions, these sanctions nonetheless constitute forms of punishment. Given that one of the main reasons for philosophers to be interested in blame is the relationship between blame and punishment, we should allow that these attitudes are forms of blame.

A second advantage of accepting Blame Attitudes and rejecting Deserved Sanctions is that, without the features described in Deserved Sanctions, the attitudes described by Blame Attitudes do not involve any commitment to strong retributive principles like this one.

⁶⁹ Joel Feinberg, "The Expressive Function of Punishment", *The Monist*, 1965.

Positive Retributivism: The fact that an agent A is blameworthy for an act or omission X constitutes a non-instrumental reason that counts in favor of sanctioning A for X.

Deserved Sanctions states that blame involves taking an agent's behavior to constitute a non-instrumental reason that counts in favor of sanctioning her. By Conative Conditions, this implies that an agent is blameworthy only if her behavior really does constitute such a reason. Thus, attitudes that conform to Deserved Sanctions involve a tacit acceptance of Positive Retributivism. Many philosophers believe that Positive Retributivism is false but do not think that every instance of blame is necessarily inappropriate. This is a credible position that should not be ruled out as simply as Deserved Sanctions would do.

One of the most interesting implications of Blame Attitudes is that there are, on this view, plausibly just two fundamental conditions of blameworthiness. First, an agent must behave in a way that expresses a morally objectionable quality of will. Second, there must be no backward-looking moral principles that prohibit sanctioning the agent for such behavior. These conditions are grounded by the features of blame given in Blame Attitudes, together with Cognitive Conditions and Conative Conditions respectively.

Because Blame Attitudes suggests that these are the only two fundamental conditions of blameworthiness, it further suggests that if any condition constitutes an excuse that renders an agent blameless for behavior despite the fact that said behavior expressed a morally objectionable quality of will, the only explanation can be that said condition makes it the case that the agent does not merit sanctions: Given that there are only two fundamental conditions of blameworthiness, and since we have assumed that the behavior in question expresses a morally objectionable quality of will, this is the only remaining explanation. Consider, for instance, the following case.

Playground: During playtime at her kindergarten, Paige finds Qian hiding under a slide and crying. Paige forms a desire to be cruel to Qian, a desire that she does not have the rational capacity to question or resist. Paige walks up to Qian and screams into Qian's face at the top of her lungs and with an anger sure to make everyone in the class take notice, "You're pathetic!"

Like many of the cruel actions performed by children, Paige's scream expresses a terribly objectionable quality of will. However, it is widely held that the fact that Paige is incapable of exercising self-control over her behavior constitutes an excuse.⁷⁰

Control (Blame): It is inappropriate to blame an agent for engaging in behavior that she lacks the rational capacity to control.

Given that agents can, like Paige, express a morally objectionable quality of will while lacking the relevant capacity for control, Blame Attitudes suggests that if Control (Blame) is true, it is because an agent who engages in behavior that she lacks the rational capacity to control does not merit sanctions for that behavior.

Control (Sanctions): An agent does not merit sanctions for behavior that she lacks the rational capacity to control.

⁷⁰ Similar incapacities, and therefore similar excuses, are thought to obtain in cases where agents act badly under the influence of a disorder, addiction, duress, etc.

Normally, the explanation is assumed to go in the opposite direction, with Control (Blame) explaining the truth of Control (Sanctions). That Blame Attitudes implies the opposite is an interesting result.⁷¹

This brings into focus one more advantage of accepting Blame Attitudes and rejecting Deserved Sanctions, namely, what this implies for the debate over whether causal determinism has any bearing on the appropriateness of moral blame. *Incompatibilists* defend the following principle.

Determinism (Blame): It is inappropriate to blame an agent for engaging in behavior in which it would be physically impossible for her to refrain from engaging.

Clearly it is possible for agents to express a morally objectionable quality of will in a deterministic universe. If Blame Attitudes is true, then, the only available explanation for Determinism (Blame) is a more fundamental claim about how determinism bears on whether an agent merits sanctions.

Determinism (Sanctions): An agent does not merit sanctions for engaging in behavior in which it would be physically impossible for her to refrain from engaging.

On my view, incompatibilist intuitions are partly driven by a sense that, if an agent was causally determined to engage in all of the behavior in which she engages, she cannot possibly merit sanctions for her behavior. Perhaps this is true; perhaps an agent only merits sanctions if she is somehow the “fundamental cause” or “origin” of her behavior. But it is more likely that this is false. According to Rawls [1955], sanctions are merited insofar as they are part of a just system

⁷¹ Rosen [2015] gives a cognitivist account of blame that has a similar result. (Gideon Rosen, “The Alethic Conception of Moral Responsibility”, *The Nature of Moral Responsibility*, 2015)

that balances the benefits of the system (namely, discouraging wrongdoing) with the costs, including the costs to those who suffer sanctions.⁷² More plausibly still, Scanlon [1986] develops a view on which sanctions are merited insofar as they are part of a system that facilitates good choices in a way that no one could reasonably reject.⁷³ If whether an agent merits sanctions depends on the kinds of standards indicated by Rawls and Scanlon and does not depend on an agent's being the fundamental cause or origin of her behavior — a position that seems to me very plausible — then Determinism (Sanctions) is likely false and so too is Determinism (Blame).

According to Deserved Sanctions, our ordinary blaming practices presuppose an esoteric connection between morally objectionable behavior and an agent's deserving sanctions. According to Blame Attitudes, it is rather the incompatibilist who is in the grip of the idea that whether sanctions are merited depends on an esoteric property like *Being fundamentally responsible for one's behavior*. The latter possibility seems the more likely. This is not to say that the incompatibilist is wrong, only that the burden is on her to show that she is right. This is an advantage of Blame Attitudes.

3.3 Objections

In this section, I consider some objections to the claims made in sections 3.1 and 3.2 that clarify the view and some of its further consequences.

Objection: To blame someone involves not just taking there to be no backward-looking prohibitions on sanctioning her, but also a desire to sanction her. This is typical of resentment and indignation, and it is possible that guilt too involves such a desire, a

⁷² John Rawls, "Two Concepts of Rules", *The Philosophical Review*, 1955.

⁷³ T.M. Scanlon, "The Significance of Choice", *The Tanner Lectures on Human Values*, 1986.

desire that is typically satisfied by the pain of guilt, which is itself a kind of psychic punishment.

The objection suggests that my earlier claim that guilt is compatible with a thoroughgoing desire to avoid sanctions is mistaken. According to the objection, guilt, like all blame attitudes, involves a desire that the offending agent (in the case of guilt, oneself) be sanctioned; what is unique about guilt, on this view, is that the pain it involves constitutes a form of sanction and can thus satisfy that desire. The fact that guilt seems compatible with a thoroughgoing desire to avoid sanctions is reinterpreted, on this view, as a desire to avoid further sanctions, to avoid suffering any sanctions other than the pain of guilt.

It seems to me that this view makes guilt too self-regarding. Guilt plausibly involves feeling bad about what one did — most likely because one has a desire not to have acted this way — but it is less plausible that it involves satisfaction at feeling this very pain. Guilt need not be so perverse. It is better, I think, to maintain that guilt does not involve any desire to receive sanctions.

As to the point about how resentment and indignation typically do involve a desire to impose sanctions, this association can be explained by the fact that if one has evidence that an agent has expressed a morally objectionable quality of will in a way that merits sanctions, then one generally has evidence that sanctions are called for, all things considered; moral sanctions are, after all, an effective form of moral education, have an important deterrent effect, etc. It will thus be rational, in many cases, for agents who feel resentment or indignation to form an instrumental desire to impose sanctions. But not in all cases. One can blame the dead, for instance — whether someone one knew or a despot whose crimes occurred long ago — without forming a desire to sanction her. It is thus a virtue of my view that blame attitudes only

conditionally involve the formation of a desire to impose sanctions, since it can account for such cases.

One could respond to this point by insisting that the desire to sanction that is typically involved in resentment and indignation is not instrumental, but intrinsic. One might have such a desire because one is in the grip of a (false) moral theory on which there are objective non-instrumental reasons to impose sanctions, or just because one finds the prospect of vengeance appealing. What evidence could there be for insisting that the desire to sanction that is typically involved in resentment and indignation is intrinsic? The only evidence I can imagine is the phenomenology of these attitudes. If resentment and indignation necessarily felt vengeful, or the suffering of the offending agent presented itself with an aura of moral goodness, this would be some evidence for the claim. On reflection, however, it seems to me that resentment and indignation need not involve such phenomenal elements. Not all resentment and indignation is vengeful. Even if some instances of resentment and indignation do involve such elements, there is good reason not to restrict our conception of blame to attitudes that involve such elements, as I argued in the previous section.

Objection: It is plausible that there are backward-looking moral principles that prohibit sanctioning agents for self-regarding behavior. If BCMS is true, then this would entail that agents are never blameworthy for self-regarding behavior. But this is false: agents can be blameworthy for self-regarding behavior, even if they do not merit sanctions for such behavior.

If this objection is to be plausible, then the self-regarding behavior in question must be behavior that is *suberogatory*, behavior that is morally bad, but not morally wrong; the fact that some wrong action — theft, say, or the murder of a rival — is in one's own self-interest does not

plausibly entail that one does not merit sanctions for that behavior. The cases that support the objection, then, are cases like this one from Julia Driver [1992].⁷⁴

[I]n boarding a train, the person who is first gets first choice of seats. Suppose that the train is almost full, and a couple wish to sit together, and there is only one place where there are two seats together. If the person ahead of them takes one of those seats, when he could have taken another less convenient seat, and knowing that the two behind him wanted to sit together, then he has done something blameworthy. (Driver [1992], 287)

Driver claims that it is both suberogatory and blameworthy for the single rider to claim one of the remaining pair of seats. One could dispute the claim that her actions are suberogatory (one might think that, although the single rider has a *prima facie* right to the convenient seat in virtue of having boarded the train first, it is nonetheless wrong for her to sit there) but let us assume that it is correct and, further, that she therefore does not merit sanctions for taking the seat. If Driver is right that the rider is nonetheless blameworthy, BCMS is false.

I deny that the rider is blameworthy. This would be implausible if it implied that agents who perform suberogatory acts deserve no moral criticism for their acts whatsoever; the rider has demonstrated a higher regard for her own convenience than for the happiness of the couple, and this is a moral fault. But there are a variety of attitudes that express moral criticism: In addition to resentment, indignation, guilt, there is also disappointment, frustration, disdain, contempt, and shame, to name a few. On my view, if it is true that the rider's actions are not

⁷⁴ Driver, Julia, "The Suberogatory", *Australasian Journal of Philosophy*, Volume 70, Number 3, 1992.

wrong, but suberogatory, then a kind of moral disappointment is appropriate, but the blame emotions of guilt, resentment, and indignation are not.⁷⁵

Objection: The way that blame is characterized in Blame Attitudes fails to account for the fact that blame is an essentially negative attitude, since it allows for there to be instances of moral blame that coexist with nothing but warm feelings and good will for the offending agent.

This objection is mistaken. My characterization of moral blame does account for the fact that moral blame is a negative attitude: For most of us, to take oneself to have no backward-looking moral reasons not to sanction someone for her behavior constitutes a serious loss of moral regard for that agent. This loss of moral regard can coexist with warm feelings and good will, but it does not do so happily. This is why, on my view, blaming one's friends and loved ones is a fraught affair, a business that becomes even more fraught when one has evidence that, all things considered, one ought to impose sanctions (say, because the friend's morally objectionable attitudes are likely to persist unless one does so). It is hard to treat someone harshly and simultaneously feel no anger, only good will. It is thus reasonable to become angry and temporarily eschew one's warm feelings in cases where sanctions are called for if it enables one to do what one should.

⁷⁵ On my considered view, this is both because the rider does not merit sanctions and because she has not expressed the kind of morally objectionable quality of will that is required for blame to be appropriate. As I argue in Chapter Two, an agent only expresses the kind of morally objectionable quality of will necessary for blameworthiness when her motives fail to counterbalance the moral reasons she has to act otherwise. Since taking the seat is suberogatory, but not wrong, the reasons of convenience that presumably motivate the rider to take the seat must then counterbalance the moral reasons she has to sit elsewhere. Although the rider's quality of will is not ideal, its badness does not, on my view, rise to the level of being morally objectionable in the sense relevant to moral blame.

Objection: The incompatibilist need not show that there is a property of fundamental desert in order to defend Determinism (Sanctions) and Determinism (Blame). For instance, if the kind framework advanced by Rawls [1955] and Scanlon [1986] supports Control (Sanctions), there is rational pressure to accept Determinism (Sanctions), too.

It is true that not every argument for Determinism (Blame) must assume that blameworthiness requires fundamental desert, but many do. For instance, Pereboom [2014] supports the position with an argument by cases. Ultimately, the challenge is for the compatibilist, who opposes Determinism (Blame), to give a principled reason for saying that an agent in one of these cases is blameless (an agent who is determined to act wrongly at a certain point because of the manipulation of his brain by a team of neuroscientists) while another agent (who is determined to act wrongly at a certain point by more normal circumstances and the laws of nature) is blameworthy. Why should it matter morally, Pereboom asks, whether the wrong action is determined by one set of microphysical events or another?

If one assumes that blameworthiness requires a kind of fundamental desert, this question can seem impossible to answer. If, however, one thinks that blameworthiness depends on whether there is a backward-looking principle no one could reasonably reject that prohibits sanctions in such cases, one can argue that while no one could reasonably reject a principle that prohibits sanctioning an agent whose behavior was determined by the intervention of another human being, we can reasonably reject a principle that prohibits sanctioning anyone whose behavior was determined by the laws of nature. Perhaps the fact that one accepts the former principle creates some kind of rational pressure to accept the latter, but not enough pressure to accept it, given such a principle's implications.

The objection, however, suggests a different argument, one that proceeds from Control (Sanctions) to Determinism (Sanctions). The argument goes thusly: when we ask whether an agent has the rational ability to refrain from a certain behavior, one takes into account features of the agent's circumstances. Consider an example of a kind of case to which Control (Sanctions) might be thought to apply.

Addiction Two: Anne is supposed to pick her daughter Belle up from school when she is struck with a craving for a narcotic a doctor prescribed her after a recent surgery. Anne knows that if she takes the narcotic, she will be unable to pick Belle up from school, leaving Belle stranded for hours. However, Anne's addictive desire for the drug is so powerful that she is incapable of resisting. Although she knows it is wrong to leave Belle at school, her powerful desire for the narcotic outweighs her desire to pick Belle up from school, rationalizing the decision to take the drug.

If Control (Sanctions) is true, then Anne arguably does not merit sanctions for failing to pick Belle up from school. It does not matter that Anne recognizes the reasons she has to pick Belle up from school and has a broad capacity to act on such reasons; rather, what matters is that she lacks the narrower capacity to pick Belle up from school given the strong desires that her addiction has produced in her on this occasion.

At this point, the incompatibilist can insist that whenever we ask whether an agent has the capacity to perform a certain act, the relevant capacity is not a general capacity to perform similar acts in similar circumstances, but the maximally-specific capacity to perform the act in question given all of the physical facts. If the agent was causally determined to not perform this

act, this is to ask whether she has the capacity to do something that is impossible. It is natural to think that the answer here must be “No”.⁷⁶

I admit that this argument creates rational pressure to accept Determinism (Sanctions) if one accepts Control (Sanctions), and does so without any need to invoke the concept of fundamental desert. As I see it, the compatibilist has three options, either (i) reject Control (Sanctions), (ii) maintain that the relevant capacity is not the maximally-specific capacity to perform the act given all of the physical facts, or (iii) maintain that an agent can have a capacity to do the impossible. Again, since I doubt that Determinism (Sanctions) is a principle that no one could reasonably reject, I am inclined to pursue one or more of these responses. In fact, I think there are independent reasons to think that (ii) and (iii) are true, but I will not pursue those arguments here.

⁷⁶ For criticism of this claim, see Lewis [1981]. (David Lewis, “Are We Free to Break the Laws?”, *Theoria*, 1981)

Conclusion

In the introduction to this essay, I introduced the concept of moral blame in terms of the paradigm blame attitudes of moral resentment, indignation, and guilt. I explained that these attitudes involve a kind of deep moral criticism, criticism both of an agent's behavior and of the agent herself. Moral blame is expressed in legal punishments and other sanctions. The main objective of this essay has been to give a more precise characterization of the attitudes of moral blame and an account of the conditions under which these attitudes are appropriate or inappropriate.

The moral criticism involved in blame is moral criticism of an agent's will, as expressed by her behavior. To blame someone, I have argued, involves taking her behavior to express a certain lack of respect for the rights and interests of others, what I call a morally objectionable quality of will. This raises two questions, addressed in Chapter Two: When does an agent's behavior express her will? and What makes the will expressed by an agent's behavior morally objectionable in the sense relevant to moral blame?

Chapter Two argued that behavior only expresses an agent's will when it is rationalized by the agent's preferences. This requirement means that behavior that is not produced by the agent's capacity for broad means-end rationality — tics, glitches, and reflexes, for instance — do not express an agent's will and therefore are not blameworthy. Recall the case of Kitten.

Kitten: Kate is across the street from a burning building where firefighters are trying to rescue a trapped kitten. They do not know what room the kitten is in and the fire is

spreading fast. Soon the firefighters will be forced to exit the building. Kate is extremely concerned for the kitten and wishes she could help.

Suddenly, Kate sees movement in a window and realizes that it's the kitten. Kate could call out to the firefighters and the kitten would be saved. All things considered, this is what Kate desires to do. However, Kate finds herself frozen with panic. A moment later, Kate loses her opportunity to act when the fire chief calls an exit from the building.

Kate's preferences support taking action to rescue the kitten. Her inaction is a kind of involuntary glitch brought on by anxiety, and thus does not express her will. This is why, although we might think Kate is a disappointment as a would-be rescuer, it would be inappropriate to be indignant or to resent her for her failure to act.

A condition that holds that agents are only blameworthy for behavior that is rationalized by their preferences also grounds an epistemic condition on blameworthiness: intuitively, one's preferences cannot rationalize risking some outcome if one has no idea that one's behavior carries such a risk. This is the ultimate explanation for the epistemic condition on blameworthiness for consequences advanced in Chapter One.

Knowledge of Risk: An agent A is blameworthy for some consequence C of behavior X only if

- (i) A believes on the basis of evidence E that X might result in a consequence of kind K, and
- (ii) C realizes K in a way such that E is not completely defeated.

Knowledge of Risk explains why agents are not blameworthy for certain outcomes that they do not suspect will obtain or that obtain as a kind of fluke.

In addition to discussing blameworthiness for consequences, Chapter One also discussed general epistemic conditions on blameworthy behavior. The chapter argued that if there is a general epistemic condition on blameworthy behavior, it plausibly takes the following form.

Restricted Explanatory Condition: The fact that an agent A is blameworthy for behavior X is always partially explained by facts of the form $E(A, M)$, where E is an epistemic attitude that A bears towards some moral reason M.

Restricted Epistemic Condition was motivated by the case of Interview, in which Irene is blameworthy for giving biased testimony to a hiring committee in virtue of her motivations for giving that testimony (i.e., dislike). Interview supports Restricted Explanatory Condition both because (i) the fact that Irene is unaware of her motives is a problem for more general theses, and (ii) Irene's awareness of the moral reasons to hire the best candidate still plausibly plays a role in explaining why she is blameworthy for giving such biased testimony.

In Chapter One, I suggested that the reason that even unconscious motives can make an agent blameworthy is that even unconscious motives contribute to the moral quality of will expressed in an agent's behavior. This brings us to the second question about quality of will raised above, namely, under what conditions the quality of will expressed in an agent's behavior is morally objectionable in the sense required for moral blame to be appropriate.

Part of what is involved in expressing a morally objectionable quality of will is an awareness of relevant moral reasons. (This is what makes Restricted Explanatory Condition true.) Chapter

Two argues that whether an agent's behavior expresses a morally objectionable quality of will depends on the relationship an agent's motivations bear to the moral reasons that she possesses and that are relevant to her behavior.

Countervailing Reasons: An agent A expresses a morally objectionable quality of will in behavior X if and only if

- (i) A's preferences rationalize Xing,
- (ii) A possesses moral reasons M to refrain from Xing, and
- (iii) A's motives for Xing do not counterbalance M.

Chapter Two argued that Countervailing Reasons gets the right result in a wide range of cases, including cases where an agent is blameworthy for behavior that is, as far as the available evidence is concerned, morally permissible, as in the case of Interview.

Chapter One suggested that epistemic conditions on blameworthiness are grounded in the condition that blameworthy behavior express a morally objectionable quality of will. Chapter Two made that notion precise, then further argued that the expression of a morally objectionable quality of will is a fundamental condition of blameworthiness, a condition grounded in the following feature of moral blame.

Characteristic Thought: If an attitude E directed at an agent A for behavior X is a way of blaming A for X, then E involves a cognitive element the content of which is true only if

- (i) A's preferences rationalize Xing,
- (ii) A possesses moral reasons M to refrain from Xing, and
- (iii) A's motives for Xing do not counterbalance M.

When we blame someone, we judge her behavior to be cruel, callous, selfish, or to otherwise involve a kind of moral vice. Characteristic Thought is a claim about what these various judgments have in common: In every case, if the criticism involved in blame is accurate, the agent meets the conditions for expressing a morally objectionable quality of will given by Countervailing Reasons. This is why Countervailing Reasons is a condition of blameworthiness.

Claims like Characteristic Thought mark a significant step forward in our theorizing about blame. We introduced the concept of moral blame by ostension, in terms of the paradigm attitudes of moral resentment, indignation, and guilt. We then noted that an agent can do something that is objectively quite bad and yet be blameless. Recall a case from the introduction.

Shiitake: Suzy is making a shiitake mushroom broth to have for dinner with Tiffany.

Unbeknownst to Suzy, Tiffany is dangerously allergic to shiitake. Suzy has never heard of a shiitake allergy, nor does Tiffany inquire into the provenance of the broth before beginning to eat. As a result, Tiffany suffers a severe allergic reaction.

Motivated by cases like Shiitake, we set out in Chapter One to determine the epistemic conditions on blameworthiness, that is, to determine how an agent's epistemic state bears on whether and for what it is appropriate to blame her. Characteristic Thought, however, brings our attention back to the nature of blame. It is called Characteristic Thought because it identifies an element that is characteristic of or essential to moral blame: a thought or judgment about the offending agent's moral will. Characteristic Thought thus supplements our ostensive definition of blame with a partial analysis of what makes an attitude like resentment, indignation, or guilt a form of moral blame.

The ambition of Chapter Three is to complete that analysis, that is, to fully characterize moral blame in terms of the attitudinal elements that it involves. The analysis I favor is as follows.

Blame Attitudes: For an agent A to blame an agent B for behavior X is for A to have an attitude E that involves:

- (i) thinking that B's performance of X has features that, as a matter of fact, are sufficient for X to express a morally objectionable quality of will, and
- (ii) taking there to be no backward-looking moral principle that prohibits sanctioning A for X.

The first condition of Blame Attitudes is a reference to Characteristic Thought. The second condition is a claim about the relationship between moral blame and moral sanctions. If Blame Attitudes is correct, then resentment, indignation, and guilt are forms of blame because they both involve the requisite thoughts about an agent's morally objectionable motives and because they involve taking there to be no restrictions of a certain kind that would otherwise prohibit sanctioning the offending agent for the behavior in question.

The prohibitions in question are backward-looking prohibitions on the impositions of sanctions, principles that prohibit sanctioning agents for behavior that has certain features or that was arrived at in a certain way. The most important of the backward-looking prohibitions that philosophers sometimes endorse are those that figure in debates over free will. For instance:

Determinism (Sanctions): An agent does not merit sanctions for engaging in behavior in which it would be physically impossible for her to refrain from engaging.

Determinism (Sanctions) is a backward-looking principle because, if it is true, the justification of moral sanctions is contingent upon a certain historical fact, a fact about the way the agent came to engage in the behavior in question and whether it was possible for her to do otherwise.

According to Blame Attitudes, blaming an agent for some behavior involves taking there to be no such backward-looking prohibitions that prohibit sanctioning her for the behavior in question. Chapter Three argues that this impacts blame's appropriateness conditions in much the same way that a characteristic thought does, as stated in the following principle.

Conative Conditions: If an attitude A involves taking a fact F to bear on whether one should Φ in virtue of norm N, then A is appropriate only if F does bear on whether one should Φ in virtue of norm N.

According to Determinism (Sanctions), there is a backward-looking moral norm that prohibits sanctioning anyone in a deterministic universe. Since blaming someone involves taking there to be no such backward-looking moral norm, Conative Conditions implies that causal determinism makes blame inappropriate. Prohibitions on sanctions like Determinism (Sanctions) thus ground corresponding conditions on the appropriateness of blame.

Determinism (Blame): It is inappropriate to blame an agent for engaging in behavior in which it would be physically impossible for her to refrain from engaging.

Incompatibilists endorse conditions like this one. Compatibilists deny them. If Blame Attitudes is true, then the more fundamental question is not whether Determinism (Blame) is true, but whether a principle like Determinism (Sanctions) is true. This is an important result for philosophers interested in free will and the ethics of moral sanctions.

What about the rest of us? What do we gain from a theory of moral blame and blameworthiness of the kind put forward in this essay? One thing we gain is an account of how a set of common moral emotions relates to substantive moral values. We blame agents for failing to be motivated by considerations that justify their actions, and for thus failing to respect the rights and interests of others. It is important that there are other kinds of moral failures and other kinds of moral emotions. The subtlety of some of the distinctions made in this essay between cases of blameless and blameworthy behavior demonstrate a need for our moral criticism to be equally subtle.

Given that criminal punishment expresses an attitude of moral blame, our theory of what makes for the expression of a morally objectionable quality of will is also an account of a kind of deep moral criticism that is involved in criminal punishment. This is of some practical significance: If an agent has not expressed a morally objectionable quality of will, then she should not be punished for that behavior. This gives us a framework for assessing the justice of criminal statutes, in particular, whether the *mens rea* or mental state necessary for conviction is sufficient for blameworthiness.

Finally, it is important that blaming someone involves taking her behavior to merit sanctions. This practical state can dispose one to treat the offending agent in ways that are unpleasant or burdensome. We should of course question whether this stance is warranted, whether we have failed to appreciate some principle that prohibits us from imposing sanctions. However, we should also remember that the license to sanction the blameworthy is conditional on forward-looking considerations. It is important to remember that we have a duty to consider what ends might be served by the imposition of sanctions, and whether this end permits — or demands — our taking such measures.

References

- Arpaly, Nomy. *Unprincipled Virtue*, Oxford University Press, 2002.
- Arpaly, Nomy, and Timothy Schroeder. *In Praise of Desire*, Oxford University Press, 2013.
- Austin, J.L. "A Plea for Excuses", *Proceedings of the Aristotelian Society*, New Series, Volume 37, 1956.
- Calhoun, Cheshire. "Changing One's Heart", *Ethics*, Volume 103, Number 1, 1992.
- D'Arms, Justin and Daniel Jacobson. "Sentiment and Value", *Ethics*, Volume 110, Number 4, 2000.
- Driver, Julia. "The Suberogatory", *Australasian Journal of Philosophy*, Volume 70, Number 3, 1992.
- Feinberg, Joel. "The Expressive Function of Punishment", *The Monist*, Volume 49, Number 3, 1965.
- Doing and Deserving*, Princeton University Press, 1970.
- Graham, Peter. "A Sketch of a Theory of Moral Blameworthiness", *Philosophy and Phenomenological Research*, 2012.
- Harman, Elizabeth. "Does Moral Ignorance Exculpate?" *Ratio*, New Series, Volume 24, 2011.
- Hieronymi, Pamela. "The Force and Fairness of Blame", *Philosophical Perspectives*, Volume 18, Number 1, 2004.
- Knobe, Joshua. *Lack of Character: Personality and Moral Behavior*, Cambridge University Press, 2002.
- Lewis, David. "Are We Free to Break the Laws?", *Theoria*, Volume 47, 1981.
- "Dispositional Theories of Value", *Proceedings of the Aristotelian Society*, Supplementary Volume 63, 1989.
- Markovits, Julia. "Acting for the Right Reasons", *The Philosophical Review*, Volume 119, Number 2, 2010.
- Milgram, Stanley. *Obedience to Authority*, Harper Perennial, 1983.
- Pereboom, Derek. *Free Will, Agency, and Meaning in Life*, Oxford University Press, 2014.
- Rawls, John. "Two Concepts of Rules", *The Philosophical Review*, Volume 64, Number 1, 1955.
- Rosen, Gideon. "Culpability and Ignorance", *Proceedings of the Aristotelian Society*, New Series, Volume 103, 2003.
- "Skepticism About Moral Responsibility", *Philosophical Perspectives*, Volume 18, Issue 1, 2004.
- "Kleinbart the Oblivious and Other Tales of Ignorance and Responsibility", *The Philosophical Review*, 2008.
- "The Alethic Conception of Moral Responsibility", *The Nature of Moral Responsibility*, Eds. Randolph Clarke, Michael McKenna, Angela M. Smith, 2015.
- Scanlon, T.M. "The Significance of Choice", *The Tanner Lectures on Human Values*, 1986.
- Moral Dimensions, Harvard University Press, 2008.
- Schroeder, Mark. *Slaves of the Passions*, Oxford University Press, 2007.
- Sher, George. *In Praise of Blame*, Oxford University Press, 2005.
- Smart, J.J.C. "Free Will, Praise and Blame", *Mind*, Volume 70, Number 279, 1961.

- Smith, Angela. "Moral Blame and Moral Protest", *Blame: It's Nature and Norms*, Eds. D. Justin Coates and Neal A. Tognazzini, 2012.
- Smith, Holly. "Culpable Ignorance", *The Philosophical Review*, Volume 92, Number 4 1983.
- Smith, Michael. "Rational Capacities or: How to Distinguish Recklessness, Weakness, and Compulsion", *Weakness of Will and Practical Irrationality*, Ed. Christine Tappolet, Oxford University Press, 2003.
- Strawson, Peter F. "Freedom and Resentment", *Proceedings of the British Academy*, 1962; reprinted in *Perspectives on Moral Responsibility*, Cornell University Press, Eds. John Martin Fischer, Mark Ravizza, 1993.
- Thomson, Judith Jarvis. *Rights, Restitution, and Risk*, Harvard University Press, 1986.
- Watson, Gary. "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme", *Responsibility, Character, and the Emotions*, Cambridge University Press, 1987; reprinted in *Agency and Answerability: Selected Essays*, Oxford University Press, 2004.
- Wallace, R. Jay. *Responsibility and the Moral Sentiments*, Harvard University Press, 1994.
- Zimmerman, Michael J. "Moral Responsibility and Ignorance", *Ethics*, Volume 107, Number 3, 1997.